

# PageRank for evolving link structures

Christopher Engström,  
Division of Applied Mathematics  
School of Education, Culture and Communication (UKK)  
Mälardalen University,  
christopher.engstrom@mdh.se

Sergei Silvestrov  
Division of Applied Mathematics  
School of Education, Culture and Communication (UKK)  
Mälardalen University  
sergei.silvestrov@mdh.se

January 24, 2014

In this article we will look at the PageRank algorithm used as part of the ranking process of different Internet pages in search engines by for example Google. This article has its main focus in the understanding of the behavior of PageRank as the system dynamically changes either by contracting or expanding such as when adding or subtracting nodes or links or groups of nodes or links. In particular we will take a look at link structures consisting of a line of nodes or a complete graph where every node links to all others.

We will look at PageRank as the solution of a linear system of equations and do our examination in both the ordinary normalized version of PageRank as well as the non-normalized version found by solving the linear system. We will see that it is possible to find explicit formulas for the PageRank in some simple link structures and using these formulas take a more in-depth look at the behavior of the ranking as the system changes.

**Keywords** PageRank Random walk Graphs Linear system

**MSC codes** 05C50 15A18 15A51 65C40

# 1 Introduction

PageRank is a method in which we can rank nodes in different link structures such as Internet pages on the Web in order of "importance" given the link structure of the complete system. It is important that the method is extremely fast since there is a huge number of Internet pages. It is also important that the algorithm returns the most relevant results first since very few people will look through more than a couple of pages when doing a search in a search engine. [7]

While PageRank was originally constructed for use in search engines, there are other uses of PageRank or similar methods, for example in the EigenTrust algorithm for reputation management to decrease distribution of unauthentic files in P2P networks. [15]

Calculating PageRank is usually done using the Power method which can be implemented very efficiently, even for very large systems. The convergence speed of the Power method and its dependence on certain parameters have been studied to some extent. For example the Power method on a graph structure such as that created by the Web will converge with a convergence rate of  $c$ , where  $c$  is one of the parameters used in the definition [12], and the problem is well conditioned unless  $c$  is very close to 1 [14]. However since the number of pages on the Web is huge, extensive work has been done in trying to improve the computation time of PageRank even further. One example is by aggregating webpages that are "close" and are expected to have a similar PageRank as in [13]. Another method used to speed up calculations is found in [1] where they do not compute the PageRank of pages that have already converged in every iteration. Other methods to speed up calculations include removing "dangling nodes" before computing PageRank and then calculate them at the end or explore other methods such as using a power series formulation of PageRank [3].

There are also work done on the large scale using PageRank and other measures in order to learn more about the Web, for example looking at the distribution of PageRank both theoretically and experimentally such as in [9].

While the theory behind PageRank is well understood from Perron-Frobenius theory for non-negative irreducible matrices [4, 11, 16] and the study of Markov chains [17, 18], how PageRank is affected from changes in the the system or parameters is not as well known.

In this article we start by giving a short introduction on PageRank and some notation and definitions used throughout the article. We will look at PageRank as the solution to a linear system of equations and what we can learn using this representation. Looking at some common graph structures we want to gain a better understanding of the changes in PageRank as the graph structure changes. This could for example be used in finding good approximations of PageRank of certain structures in order to speed up calculations further. We will look at both the "ordinary" normalized version of PageRank as well as a non-normalized version we get by solving the linear system. We will see how this non-normalized version corresponds to the probabilities of a random walk through the graph and how we can use this to find the PageRank of some systems using this perspective rather than solving the system or computing the dominant eigenvector. Mainly two different structures, first a simple line in Sect. 4.1 and later a complete graph in Sect. 4.2

will be examined. In both cases we will see that we can find explicit expressions for the PageRank depending on the number of nodes. In both cases of the "ordinary" PageRank as well as a non-normalized version expressions for the PageRank will be found for both the structure itself as well as the PageRank after doing some simple modifications. The last graph structure we will look at is when we combine the simple line with the complete graph by adding a link between them in Sect. 4.3. In Sect. 4.4 and Sect. 4.5 we will take a closer look at the found formulas for some of the examples mainly by looking at partial derivatives of the PageRank. We will see one of the possible reasons why  $c$  is usually chosen to be around  $c \approx 0.85$ . PageRank for some nodes increases extremely fast while for some other nodes decreases extremely fast for larger  $c$ , while for lower  $c$  the difference in PageRank between nodes is smaller the lower  $c$  gets and the initial weight vector have a much larger influence on the final ranking. Last we take a short look at what happens when changing the weight vector  $\mathbf{V}$  present in the PageRank formulation as well as giving a short comparison of the differences and similarities between normalized and non-normalized PageRank.

## 2 Calculating PageRank

Starting with a number of nodes (Internet pages) and the non-negative matrix  $\mathbf{A}$  with every element  $a_{ij} \neq 0$  corresponding to a link from node  $i$  to node  $j$ . The value of element  $a_{ij} = 1/n$  where  $n$  is the number of outgoing links from node  $i$ . An example of a graph and corresponding matrix can be seen in Fig. 1.

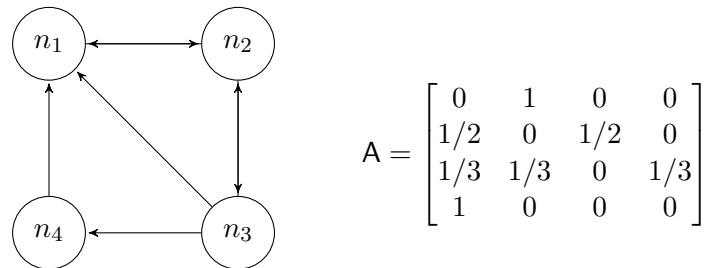


Figure 1: Directed graph and corresponding matrix system matrix  $\mathbf{A}$

Note that we by convention do not allow a node to link to itself. We also need that no nodes have zero outgoing links (dangling nodes) resulting in a row with all zeros. For now we assume that none of these dangling nodes are present in the link matrix. This means that every row will sum to one in the link matrix  $\mathbf{A}$ .

The PageRank vector  $\mathbf{R}$  we want for ranking the nodes (pages) is the eigenvector corresponding to the eigenvalue one of matrix  $\mathbf{M}$ :

$$\mathbf{M} = c\mathbf{A}^\top + (1 - c)\mathbf{u}\mathbf{e}^\top$$

where  $0 < c < 1$ , usually  $c = 0.85$ ,  $\mathbf{A}$  the link matrix,  $\mathbf{e}$  a column vector of the same length as the number of nodes ( $n$ ) filled with ones and  $\mathbf{u}$  is a column vector of the same

length with elements  $u_i$ ,  $0 \leq u_i \leq 1$  such that  $\|\mathbf{u}\|_1 = 1$ . For  $\mathbf{u}$  we will usually use the uniform vector (all elements equal) with  $u_i = 1/n$  where  $n$  is the number of nodes. The result after calculating the PageRank of the example matrix for the system in Fig. 1 can be seen below:

$$\mathbf{R} \approx \begin{bmatrix} 0.3328 \\ 0.3763 \\ 0.1974 \\ 0.0934 \end{bmatrix}$$

This can be seen as a random walk where we start in a random node depending on the weightvector  $\mathbf{u}$ . Then with a probability  $c$  we go to any of the nodes linked to from that node and with a probability  $1 - c$  we instead go to a random (in the case of uniform  $\mathbf{u}$ ) new node. The PageRank vector can be seen as the probability that you after a long time is located in the node in question.[3] More on why an eigenvector with eigenvalue 1 always exists can be seen in for example [8].

**Role of  $c$ .** Looking at the formula it is not immediately obvious why we demand  $0 < c < 1$  and what role  $c$  holds. We can easily see what happens at the limits, if  $c = 0$  the PageRank is decided only by the initial weights  $\mathbf{u}$ . However if  $c = 1$  the weights have no role and the algorithm used for calculating PageRank might not even converge. As  $c$  increases, nodes further and further away have an impact on the PageRank of individual nodes. And the opposite for low  $c$ , the lower  $c$  is the more important is the immediate surrounding of a node in deciding its PageRank. The parameter  $c$  is also a very important factor in how fast the algorithms used to calculate PageRank converges, the higher  $c$  is the slower the algorithm will converge.

**Handling of dangling nodes.** If  $A$  contains dangling nodes, corresponding row no longer sums to one and there therefor will probably not be any eigenvector with eigenvalue equal to one. The method we use in order to fix this is to instead assume that the dangling nodes link to all nodes equally (or according to some other desired distribution). This gives us:  $\mathbf{T} = \mathbf{A} + \mathbf{g}\mathbf{w}^\top$ , where  $\mathbf{g}$  is a column vector with elements equal to one for a dangling node and zero for all other nodes. Here  $\mathbf{w}$  is the distribution according to how we make the dangling nodes link to other nodes (usually uniform or equal to  $\mathbf{u}$ ). In this work we always use  $\mathbf{w} = \mathbf{u}$  to simplify calculations.

There are other ways to handle dangling nodes, for example by adding one new node linking only to itself and let all dangling nodes link to this node. Assuming  $\mathbf{w} = \mathbf{u}$  these methods should be essentially the same apart from implementation [6].

### 3 Notation and definitions

Here we give some notes on the notation used through the rest of the article in order to clarify which variation of PageRank is used as well as some overall notation and the definition of some common important link structures. We will repeatedly use the  $L^1$  norm in comparing the size of different vectors or (parts of) matrices.

First some overall notation:

- $S_G$ : The system of nodes and links for which we want to calculate PageRank, contains the system matrix  $A_G$  as well as a weight vector  $\mathbf{v}_G$ . Subindex  $G$  can be either a capital letter or a number in the case of multiple systems.
- $n_G$ : The number of nodes in system  $S_G$ .
- $A_G$ : System matrix where a zero element  $a_{ij}$  means there is no link from node  $i$  to node  $j$ . Non-zero elements are equal to  $1/r_i$  where  $r_i$  is the number of links from node  $i$ . Size  $n_G \times n_G$ .
- $\mathbf{v}_G$ : Non-negative weight vector, not necessary with sum one. Size  $n_G \times 1$ .
- $\mathbf{u}_G$ : The weight vector  $\mathbf{v}_G$  normalized such that  $\|\mathbf{u}_G\|_1 = 1$ . We note that  $\mathbf{u}_G$  is proportional to  $\mathbf{v}_G$  ( $\mathbf{u}_G \propto \mathbf{v}_G$ ). Size  $n_G \times 1$ .
- $c$ : Parameter  $0 < c < 1$  for calculating PageRank, usually  $c = 0.85$ .
- $\mathbf{g}_G$ : Vector with elements equal to one for dangling nodes and zero for all other in  $S_G$ . Size  $n_G \times 1$ .
- $M_G$ : Modified system matrix,  $M_G = c(A_G + \mathbf{g}_G \mathbf{u}_G^\top)^\top + (1-c)\mathbf{u}_G \mathbf{e}^\top$  used to calculate PageRank, where  $\mathbf{e}$  is the unit vector. Size  $n_G \times n_G$ .
- $S$ : Global system made up of multiple disjoint subsystems  $S = S_1 \cup S_2 \dots \cup S_N$ , where  $N$  is the number of subsystems.
- $\mathbf{V}$ : Global weight vector for system  $S$ ,  $\mathbf{V} = [\mathbf{v}_1^\top \ \mathbf{v}_2^\top \ \dots \ \mathbf{v}_N^\top]^\top$ , where  $N$  is the number of subsystems.

In the cases where there is only one possible system the subindex  $G$  will often be omitted. For the systems making up  $S$  we define disjoint systems in the following way.

**Definition 3.1.** *Two systems  $S_1, S_2$  are disjoint if there are no paths from any nodes in  $S_1$  to  $S_2$  or from any nodes in  $S_2$  to  $S_1$ .*

From earlier we saw how we could calculate PageRank for a system  $S$ , now we make the assumption that  $\mathbf{w} = \mathbf{u}$  both since it simplifies calculations, but also since using two different weight vectors for essentially the same thing seems like it could create more problems and unexpected behavior than what you actually could gain from it.

We will use three different ways to define the different versions of PageRank using the notation:

$$\mathbf{R}_G^{(t)}[S_H \rightarrow S_I, S_J \rightarrow S_K \dots]$$

where  $t$  is the type of PageRank used,  $S_G \subseteq S$  is the nodes in the global system  $S$  for which  $\mathbf{R}$  is the PageRank. Often  $S_G = S$  and we write it as  $\mathbf{R}_S^{(t)}$ . In the last part within brackets we write possible connections between otherwise disjoint subsystems in  $S$ , for example an arrow to the right means there are links from the left system to the

the right system. How many and what type of links however needs to be specified for every individual case. In more complicated examples there may be arrows pointing in two directions or a number above the arrow notifying how many links we have between the systems.

We will sometimes give the formula for a specific node  $j$  in this case it will be noted as  $\mathbf{R}_{G,j}^{(t)}[S_H \rightarrow S_I, S_J \rightarrow S_K \dots]$ . When it is obvious which system to use (for example when only one is specified) and there are no connections between systems  $S_G$  as well as the brackets with connections between systems will usually be omitted resulting in  $\mathbf{R}_j^{(t)}$ . It should be obvious when this is the case. When normalizing the resulting elements such that their sum equal to one we get the traditional PageRank:

**Definition 3.2.**  $\mathbf{R}_G^{(1)}$  for system  $S_G$  is defined as the eigenvector with eigenvalue one to the matrix  $\mathbf{M}_G = c(\mathbf{A}_G + \mathbf{g}_G \mathbf{u}_G^\top)^\top + (1 - c)\mathbf{u}_G \mathbf{e}^\top$ .

Note that we always have  $\|\mathbf{R}^{(1)}\|_1 = 1$  and that non-zero elements in  $\mathbf{R}_G^{(1)}$  are all positive. The fact that  $\|\mathbf{R}^{(1)}\|_1 = 1$  is generally not the case in our other versions of PageRank. When instead setting up the resulting equation system and solving it we get the second definition, the result is multiplied with  $n_G$  in order to get multiplication with the one vector in case of uniform  $\mathbf{u}_G$ .

**Definition 3.3.**  $\mathbf{R}_G^{(2)}$  for system  $S_G$  is defined as  $\mathbf{R}_G^{(2)} = (\mathbf{I} - c\mathbf{A}_G^\top)^{-1}n_G\mathbf{u}_G$

We note that generally  $\|\mathbf{R}^{(2)}\|_1 \neq 1$  as well as  $\mathbf{R}_G^{(2)} \neq n_G\mathbf{R}_G^{(1)}$  unless there are no dangling nodes in the system. However the two versions of PageRank are proportional to each other ( $\mathbf{R}_G^{(2)} \propto \mathbf{R}_G^{(1)}$ ). Last we have the third way to define PageRank which we define in order to make it possible to use the power method but still be able to compare PageRank between different subsystems  $S_G, S_H, \dots$  without any additional computations as well as simplifying the work when updating the system.

**Definition 3.4.**  $\mathbf{R}_G^{(3)}$  for system  $G$  is defined as:

$$\mathbf{R}_G^{(3)} = \frac{\mathbf{R}_G^{(1)} \|\mathbf{v}_G\|_1}{d_G}$$

$$d_G = 1 - \sum c\mathbf{A}_G^\top \mathbf{R}_G^{(1)}$$

where  $\mathbf{v}_G$  is the part of the global weight vector  $\mathbf{V}$  belonging to the nodes in system  $S_G$  [10].

A closer look at  $\mathbf{R}_G^{(3)}$  is left for a later article. The definition of  $\mathbf{R}_G^{(3)}$  is included here only for completeness.

**Definition 3.5.** A simple line is a graph with  $n_L$  nodes where node  $n_L$  links to node  $n_{L-1}$  which in turn links to node  $n_{L-2}$  all the way until node  $n_2$  link to node  $n_1$ .

The link matrix  $\mathbf{A}_L$  and graph for system  $S_L$  consisting of a simple line with 5 nodes can be seen in Fig. 2:

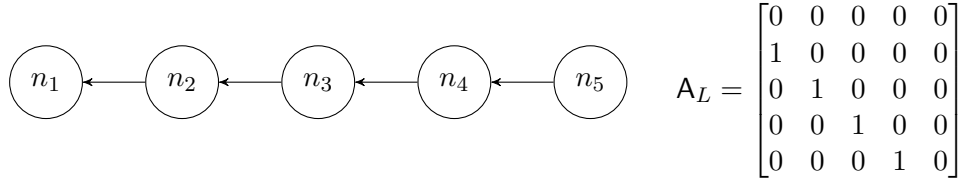


Figure 2: The simple line with 5 nodes and corresponding system matrix

**Definition 3.6.** A complete graph is a group of nodes in which all nodes in the group links to all other nodes in the group.

The link matrix  $A_G$  for system  $S_G$  consisting of a complete graph with 5 nodes can be seen in Fig. 3

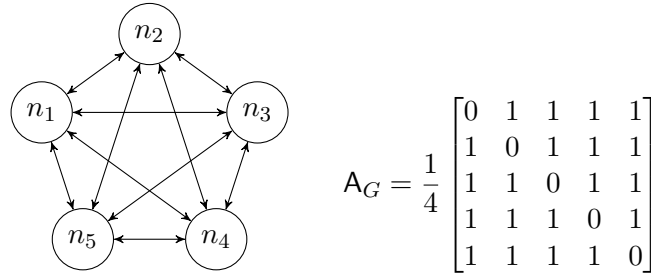


Figure 3: A complete graph with five nodes and corresponding system matrix

## 4 Changes in PageRank when modifying some common structures

Looking at some simple structures and how PageRank changes as we change them, the goal is to learn something in how and why the rank changes as it does. This in an attempt to answer questions such as: How do I connect my two sites or within my one site in such a way that I won't get any undesired results? In all these examples we will assume uniform  $\mathbf{u}$  (which means we can multiply the inverse  $(I - cA^\top)^{-1}$  with the one vector in order to get  $\mathbf{R}^{(2)}$ ).

In this article we will look at two methods two calculate PageRank ( $\mathbf{R}^{(2)}$ ), while maybe not as useful for calculating PageRank for large systems we use these ways in hope that we can learn something about the behavior of different typical systems or structures within a system. From earlier we have:

$$\mathbf{R}^{(1)} = M\mathbf{R}^{(1)} = (c(A + \mathbf{g}\mathbf{u}^\top)^\top + (1 - c)\mathbf{u}\mathbf{e}^\top)\mathbf{R}^{(1)} \quad (1)$$

Calculating the dominant eigenvector  $\mathbf{R}^{(1)}$  is the same as solving the linear system:

$$\mathbf{R}^{(1)} = \mathbf{M}\mathbf{R}^{(1)} \Leftrightarrow (c\mathbf{A}^\top - \mathbf{I})\mathbf{R}^{(1)} = -(c\mathbf{u}\mathbf{g}^\top + (1-c)\mathbf{u}\mathbf{e}^\top)\mathbf{R}^{(1)} \quad (2)$$

Since every column of  $\mathbf{u}\mathbf{g}^\top$  is either equal to  $\mathbf{u}$  or zero and all columns equal to  $\mathbf{u}$  for  $\mathbf{u}\mathbf{e}^\top$  we can see that  $-(c\mathbf{u}\mathbf{g}^\top + (1-c)\mathbf{u}\mathbf{e}^\top)\mathbf{R}^{(1)}$  will be proportional to  $\mathbf{u}$ . This can be written as:  $(c\mathbf{A}^\top - \mathbf{I})\mathbf{R}^{(1)} = k\mathbf{u}$ .

We choose  $k = -n$  in order to get  $k\mathbf{u}$  equal to the one vector in the case of uniform  $\mathbf{u}$ , the minus sign to get positive rank and solving the system we get:

$$\mathbf{R}^{(2)} = (\mathbf{I} - c\mathbf{A}^\top)^{-1}n\mathbf{u} \quad (3)$$

To get the rank to sum to one it is a simple matter of normalizing the result.  $\mathbf{R}^{(1)} = \mathbf{R}^{(2)} / \|\mathbf{R}^{(2)}\|_1$  [6]. We note the similarity with this formulation of PageRank (solution to  $\mathbf{R}^{(2)} = c\mathbf{A}^\top\mathbf{R}^{(2)} + n\mathbf{u}$ ) with the one for the potential of a Markov chain with a discounted cost (solution to  $\mathbf{R}^{(2)} = \alpha\mathbf{A}\mathbf{R}^{(2)} + c$ ), where  $0 < \alpha < 1$  is the discount factor and  $c$  is a cost vector. [17]

Note that we do not need to take any care of the dangling nodes when calculating the PageRank in this way although it is a lot slower than using the Power method or other conventional methods of calculating PageRank. Although we do not need to change  $\mathbf{A}$  for dangling nodes, the result when doing so is changed (but still proportional to  $\mathbf{R}^{(1)}$ ). We will never change  $\mathbf{A}$  for dangling nodes when solving the linear system and only use the version defined above. Note that while solving the equation system is slow it could be possible to get to this non-normalized version of PageRank using another PageRank algorithm, such as using a power series formulation as in [2].

The following theorem explains how PageRank ( $\mathbf{R}^{(2)}$ ) can be computed and how it can be interpret from a probabilistic viewpoint using random random walks on a graph and hitting probabilities.

**Theorem 4.1.** *Consider a random walk on a graph described by  $c\mathbf{A}$  described as before. We walk to a new node with probability  $c$  and stop with probability  $1 - c$ .*

*PageRank  $\mathbf{R}^{(2)}$  of a node when using uniform  $\mathbf{u}$  can be written:*

$$\mathbf{R}_j^{(2)} = \left( \sum_{e_i \in S, e_i \neq e_j} P(e_i \rightarrow e_j) + 1 \right) \left( \sum_{k=0}^{\infty} (P(e_j \rightarrow e_j))^k \right) \quad (4)$$

where  $P(e_i \rightarrow e_j)$  is the probability to hit node  $e_j$  in a random walk starting in node  $e_i$  described as above. This can be seen as the expected number of visits to  $e_j$  if we do multiple random walks, starting in every node once.

*Proof.*  $(c\mathbf{A}^\top)_{ij}^k$  is the probability to be in node  $e_i$  starting in node  $e_j$  after  $k$  steps. Multiplying with the unit vector  $\mathbf{e}$  (vector with all elements equal to one) therefor gives the sum of all the probabilities to be in node  $e_i$  after  $k$  steps starting in every node once. The expected total number of visits is the sum of all probabilities to be in node  $e_i$  for every step starting in every node:

$$\mathbf{R}_j^{(2)} = \left( \left( \sum_{k=0}^{\infty} (c\mathbf{A}^\top)^k \right) \mathbf{e} \right)_j \quad (5)$$



$\sum_{k=0}^{\infty} (c\mathbf{A}^\top)^k$  is the Neumann series of  $(\mathbf{I} - c\mathbf{A}^\top)^{-1}$  which is guaranteed to converge since  $c\mathbf{A}^\top$  is non-negative and have column sum  $< 1$ . If  $\mathbf{u}$  is uniform we get by the definition:

$$\begin{aligned}\mathbf{R}^{(2)} &= (\mathbf{I} - c\mathbf{A}^\top)^{-1} n\mathbf{u} = (\mathbf{I} - c\mathbf{A}^\top)^{-1} \mathbf{e} = \left( \sum_{k=0}^{\infty} (c\mathbf{A}^\top)^k \right) \mathbf{e} \\ \Rightarrow \mathbf{R}_j^2 &= \left( \sum_{e_i \in S, e_i \neq e_j} P(e_i \rightarrow e_j) + 1 \right) \left( \sum_{k=0}^{\infty} (P(e_j \rightarrow e_j))^k \right)\end{aligned}\tag{6}$$

□

□

#### 4.1 Changes in the simple line

Using the simple line as defined earlier we recall that we had the link matrix with an image of the system in Fig. 2

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

By setting up the system of equations we get the inverse  $(\mathbf{I} - c\mathbf{A}^\top)^{-1}$  as:

$$(\mathbf{I} - c\mathbf{A}^\top)^{-1} = \begin{bmatrix} 1 & c & c^2 & c^3 & c^4 \\ 0 & 1 & c & c^2 & c^3 \\ 0 & 0 & 1 & c & c^2 \\ 0 & 0 & 0 & 1 & c \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Note that this needs only to be multiplied with  $n\mathbf{u}$  or a multiple of  $\mathbf{u}$  for us to get a meaningful ranking. This gives us  $\mathbf{R}^{(2)}$  (for uniform  $\mathbf{u}$ ):

$$\mathbf{R}^{(2)} = [1 + c + c^2 + c^3 + c^4, 1 + c + c^2 + c^3, 1 + c + c^2, 1 + c, 1]^\top$$

If wanted to get the common normalized ranking  $\mathbf{R}^{(1)}$  we need to normalize the result to sum to one. Looking at the elements  $a_{ij}$  of  $(\mathbf{I} - c\mathbf{A}^\top)^{-1}$  and considering the example with a random walk through the graph, we can see the value of every element  $a_{ij}$  as the probability to get from node  $e_j$  to node  $e_i$ . In the case where the link matrix contain nodes with paths back to itself we will later see that it is actually not the probability to get there but the sum of all probabilities to get from  $e_j$  to  $e_i$  corresponding to Theorem 4.1. We can motivate this further by looking at the same line but adding a link back from the first node to the second node.

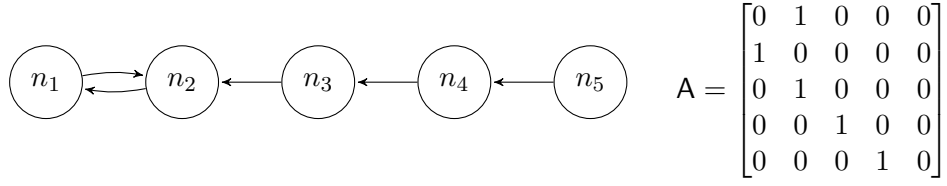


Figure 4: Simple line where the first node links to the second and corresponding system matrix

#### 4.1.1 The simple line with node one linking to node two

Letting node one link to node two in the earlier example gives us the graph in Fig. 4. The resulting inverse can be written:

$$(\mathbf{I} - c\mathbf{A}^\top)^{-1} = \begin{bmatrix} s & sc & sc^2 & sc^3 & sc^4 \\ sc & s & sc & sc^2 & sc^3 \\ 0 & 0 & 1 & c & c^2 \\ 0 & 0 & 0 & 1 & c \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Where  $s = \sum_{k=0}^{\infty} c^{2k} = \frac{1}{1-c^2}$  is the sum of all the probabilities of getting from node 1 or 2 back to itself.

From this we can see that the following observations seem to be true.

- (i) The sum of a column  $c_j$  is at most  $\sum_{k=0}^{\infty} c^k = \frac{1}{1-c}$  when using uniform  $\mathbf{u}$ , with equality if there are no paths to any dangling node from node  $j$  and node  $j$  is not a dangling node itself.
- (ii) A diagonal element is equal to one if the node have no paths leading back to itself.
- (iii) Setting one element in  $u_i$  to zero only effects the influence of a random walk starting in the corresponding node.
- (iiii) Every non zero element in the same row can be written as the diagonal element on the same line times the sum of probabilities of getting from all other nodes to the node corresponding to the current line.
- (iiiii) Each element  $e_{ij}$  of  $(\mathbf{I} - c\mathbf{A}^\top)^{-1}$  contains the sum of probabilities of all paths starting in node  $j$  and ending in node  $i$ . When doing a random walk by choosing a random link with probability  $c$  and stopping with probability  $1 - c$ .

Which is consistent with the statement that the normalized PageRank  $\mathbf{R}_j^{(1)}$  of a node is the probability that a surfer that starts in a random node (page) and keeps clicking links with probability  $c$  or starts at a new random page with probability  $(1-c)$  is in a given node. However here we can explicitly see all the probabilities and their influence on the ranking. [8]

#### 4.1.2 Removing a link between two nodes

When removing a link between two nodes in the simple line we end up with two smaller disjoint lines instead. We note that these could be calculated separately and we would still have the same relation between them. This is interesting since when using the "Power method" or straight calculating  $\mathbf{R}^{(1)}$  this is not possible since more nodes in a system obviously means a lower mean rank since we in that case normalize the result to one.

Especially in the inverse  $(\mathbf{I} - c\mathbf{A}^\top)^{-1}$  we see that when we remove one link, we remove all the elements in the upper right corresponding to paths from nodes above the removed link to all the ones below it. An example of what the new inverse looks like when removing the link between the third node and the second node in Fig. 2 can be seen below:

$$(\mathbf{I} - c\mathbf{A}^\top)^{-1} = \begin{bmatrix} 1 & c & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & c & c^2 \\ 0 & 0 & 0 & 1 & c \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

With PageRank:  $\mathbf{R}^{(2)} = [1 + c, 1, 1 + c + c^2, 1 + c, 1]$  and normalizing constant  $N = 5 + 3c + c^2$ , when using a uniform  $\mathbf{u}$

#### 4.1.3 Adding a new node pointing at one node in the simple line

A more interesting example is when looking at what happens when we add a single new node, linking to one other node in the simple line. Since we make no changes in the line that part of the inverse will stay the same. We will however add a new row and column. The non diagonal element of the new column can be found immediately as  $c$  times the column corresponding to the node our new node links to. This since we got the probability  $c$  to get to that node instead of 1 when we start in it. At last we need to add the one at the new element in the diagonal. An example of what the inverse looks like after adding a new node pointing at node 3 in Fig. 2 can be seen below.

$$(\mathbf{I} - c\mathbf{A}^\top)^{-1} = \begin{bmatrix} 1 & c & c^2 & c^3 & c^4 & c^3 \\ 0 & 1 & c & c^2 & c^3 & c^2 \\ 0 & 0 & 1 & c & c^2 & c \\ 0 & 0 & 0 & 1 & c & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

From this easy example we can immediately get an expression for the PageRank of a simple line with one or more added nodes linking to any of the nodes in the simple line.

**Theorem 4.2.** *The PageRank of a node  $e_i$  belonging to the line in a system containing a simple line with one outside node linking to one of the nodes in the line when using uniform weight vector  $\mathbf{u}$  can be written:*

$$\begin{aligned} \mathbf{R}_i^{(2)} &= \sum_{k=0}^{n_L-i} c^k + b_{ij} = \frac{1 - c^{n_L-i+1}}{1 - c} + b_{ij} \\ b_{ij} &= \begin{cases} c^{j+1-i}, & j \geq i \\ 0, & j < i \end{cases} \end{aligned} \quad (7)$$

where  $n_L$  is the number of nodes in the line and the new node link to node  $j$ . The new node has rank 1. After normalization we get the PageRank of node  $i$  as:

$$\mathbf{R}_i^{(1)} = \frac{\frac{1 - c^{n_L-i+1}}{1 - c} + b_{ij}}{n_L + 1 + (n_L - 1)c + (n_L - 2)c^2 + \dots + c^{n_L-1} + \frac{1 - c^j}{1 - c}} \quad (8)$$

where  $\mathbf{R}_i^{(1)}, \mathbf{R}_i^{(2)}$  is the PageRank of one of the nodes in the original line,  $L$  is the number of nodes on the line,  $j$  is the number of the node linked to be the new node.

Additionally adding new nodes linking to the line means adding additional  $b_{ij}$  parts and adding the corresponding part  $\frac{1 - c^j}{1 - c}$  to the normalizing constant.

*Proof.* From Theorem 4.1 PageRank for a node when using uniform  $\mathbf{u}$  can be written as:

$$\mathbf{R}_i^{(2)} = \left( \sum_{e_j \in S, e_j \neq e_i} P(e_j \rightarrow e_i) + 1 \right) \left( \sum_{k=0}^{\infty} (P(e_i \rightarrow e_i))^k \right)$$

where  $P(e_j \rightarrow e_i)$  is the probability to hit node  $e_i$  starting in node  $e_j$ . When we consider a random walk on a graph given by  $cA$  described as before. We walk to a new node with probability  $c$  and stop with probability  $1 - c$ .

The probability of getting to any node  $e_i$  in the line from any other node  $e_j$  in the line once is:

$$P(e_j \rightarrow e_i) = c^{j-i}, \quad j > i \quad (9)$$

and zero otherwise. Summation over all  $j > i$  gives

$$\sum_{e_j \in S, e_j \neq e_i} P(e_j \rightarrow e_i) + 1 = \sum_{k=1}^{n_L-i} c^k + 1 = \frac{1 - c^{n_L-i+1}}{1 - c} \quad (10)$$

where  $L$  is the number of nodes in the line. With the first part shown we need to show that the single outside node linking to node  $e_j$  adds  $b_{ij} = c^{j+1-i}$ ,  $j \geq i$ . We get this probability in the same way by instead looking at the line created by the first  $j$  nodes plus the extra node added linking to node  $j$ . We get the probability to reach node  $e_j$  as

$c$  and then  $c^2$  for the next and so on. If  $i > j$ ,  $e_i$  does not belong to this line, and we obviously cannot reach it from  $e_j$  hence  $b_{ij} = 0$ ,  $i > j$ .

Last the PageRank of the "outside" node linking to a node in the line is obviously 1 since no node links to it. The normalized PageRank is found by dividing  $\mathbf{R}^{(2)}$  with  $\|\mathbf{R}^{(2)}\|_1$ .  $\square$   $\square$

We also give a proof using matrices but first we will need the following lemma for blockwise inversion used repeatedly throughout the article.. We note that we label the blocks from B to E rather than from A to D in order to avoid confusion with the system matrix A.

**Lemma 4.1.**

$$\begin{bmatrix} B & C \\ D & E \end{bmatrix}^{-1} = \begin{bmatrix} (B - CE^{-1}D)^{-1} & -(B - CE^{-1}D)^{-1}CE^{-1} \\ -E^{-1}D(B - CE^{-1}D)^{-1} & E^{-1} + E^{-1}D(B - CE^{-1}D)^{-1}CE^{-1} \end{bmatrix} \quad (11)$$

where B, E is square and E,  $(B - CE^{-1}D)$  are nonsingular.

*Proof.* To prove the Lemma it is enough to show that:

$$\begin{bmatrix} B & C \\ D & E \end{bmatrix} \begin{bmatrix} (B - CE^{-1}D)^{-1} & -(B - CE^{-1}D)^{-1}CE^{-1} \\ -E^{-1}D(B - CE^{-1}D)^{-1} & E^{-1} + E^{-1}D(B - CE^{-1}D)^{-1}CE^{-1} \end{bmatrix} = I \quad (12)$$

Looking at the result blockwise we get:

$$\begin{aligned} & B(B - CE^{-1}D)^{-1} - CE^{-1}D(B - CE^{-1}D)^{-1} \\ & = (B - CE^{-1}D)(B - CE^{-1}D)^{-1} = I \end{aligned} \quad (13)$$

$$\begin{aligned} & -B(B - CE^{-1}D)^{-1}CE^{-1} + C(E^{-1} + E^{-1}D(B - CE^{-1}D)^{-1}CE^{-1}) \\ & = CE^{-1} - (B - CE^{-1}D)(B - CE^{-1}D)^{-1}CE^{-1} = CE^{-1} - ICE^{-1} = 0 \end{aligned} \quad (14)$$

$$\begin{aligned} & D(B - CE^{-1}D)^{-1} - EE^{-1}D(B - CE^{-1}D)^{-1} \\ & = D(B - CE^{-1}D)^{-1} - D(B - CE^{-1}D)^{-1} = 0 \end{aligned} \quad (15)$$

$$\begin{aligned} & -D(B - CE^{-1}D)^{-1}CE^{-1} + E(E^{-1} + E^{-1}D(B - CE^{-1}D)^{-1}CE^{-1}) \\ & = -D(B - CE^{-1}D)^{-1}CE^{-1} + I + D(B - CE^{-1}D)^{-1}CE^{-1} = I \end{aligned} \quad (16)$$

This gives:

$$\begin{aligned} \begin{bmatrix} B & C \\ D & E \end{bmatrix} \begin{bmatrix} (B - CE^{-1}D)^{-1} & -(B - CE^{-1}D)^{-1}CE^{-1} \\ -E^{-1}D(B - CE^{-1}D)^{-1} & E^{-1} + E^{-1}D(B - CE^{-1}D)^{-1}CE^{-1} \end{bmatrix} \\ = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} = I \end{aligned} \quad (17)$$

Furthermore we need that E and  $(B - CE^{-1}D)$  is nonsingular in order for the matrix to be invertible. [5]  $\square$   $\square$

When using Lemma 4.1 we will denote the individual blocks of the inverse matrix as described in Definition 4.1

**Definition 4.1.** Given a block matrix  $\mathbf{M}$  we denote the inverse as:

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_{1,1} & \mathbf{M}_{1,2} & \dots & \mathbf{M}_{1,n} \\ \mathbf{M}_{2,1} & \mathbf{M}_{2,2} & \dots & \mathbf{M}_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{M}_{m,1} & \mathbf{M}_{m,2} & \dots & \mathbf{M}_{m,n} \end{bmatrix}, \quad \mathbf{M}^{-1} = \begin{bmatrix} \mathbf{M}_{1,1}^{inv} & \mathbf{M}_{1,2}^{inv} & \dots & \mathbf{M}_{1,n}^{inv} \\ \mathbf{M}_{2,1}^{inv} & \mathbf{M}_{2,2}^{inv} & \dots & \mathbf{M}_{2,n}^{inv} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{M}_{m,1}^{inv} & \mathbf{M}_{m,2}^{inv} & \dots & \mathbf{M}_{m,n}^{inv} \end{bmatrix} \quad (18)$$

We can now give a matrix proof of Theorem 4.2 as well.

*Proof of Theorem 4.2.* We let  $\mathbf{B}$  be the part of the matrix  $(\mathbf{I} - c\mathbf{A}^\top)$  corresponding to the nodes in the line which gives:

$$(\mathbf{I} - c\mathbf{A}^\top) = \begin{bmatrix} \mathbf{B} & \mathbf{C} \\ \mathbf{0} & \mathbf{1} \end{bmatrix} \quad (19)$$

We write

$$(\mathbf{I} - c\mathbf{A}^\top)^{-1} = \begin{bmatrix} \mathbf{B}^{inv} & \mathbf{C}^{inv} \\ \mathbf{D}^{inv} & \mathbf{E}^{inv} \end{bmatrix} \quad (20)$$

Using Lemma 4.1 for blockwise inverse we get  $\mathbf{B}^{inv} = (\mathbf{B} - \mathbf{C}\mathbf{E}^{-1}\mathbf{D})^{-1} = \mathbf{B}^{-1}$ . Since  $\mathbf{B}$  is the matrix for the simple line found earlier we get:

$$\mathbf{B}^{inv} = (\mathbf{I} - c\mathbf{A}^\top)^{-1} = \begin{bmatrix} 1 & c & c^2 & \dots & c^{L-1} \\ 0 & 1 & c & \dots & c^{L-2} \\ 0 & 0 & 1 & \dots & c^{L-3} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & 0 & 1 \end{bmatrix} \quad (21)$$

where  $L$  is the total number of nodes in the line.  $\mathbf{C} = [0 \dots c \ 0 \dots 0]^\top$  where the non-zero element  $c$  is at position  $j$  gives:

$$\mathbf{C}^{inv} = -\mathbf{B}^{inv}\mathbf{C}\mathbf{E}^{-1} = -\mathbf{B}^{inv}\mathbf{C} = [c^j \ c^{j-1} \ \dots \ c \ 0 \dots 0]^\top \quad (22)$$

Last, since  $\mathbf{D} = \mathbf{0}$  we get  $\mathbf{D}^{inv} = \mathbf{0}$ ,  $\mathbf{E}^{inv} = \mathbf{1}$ . Since the weight vector  $\mathbf{u}$  is uniform we get the PageRank of a node as the sum of corresponding row in  $(\mathbf{I} - c\mathbf{A}^\top)^{-1}$ . For the nodes in the line we get PageRank:

$$\mathbf{R}_i^{(2)} = \sum_{k=0}^{n_L-i} c^k + b_{ij} = \frac{1 - c^{n_L-i+1}}{1 - c} + b_{ij} \quad (23)$$

$$b_{ij} = \begin{cases} c^{j+1-i}, & j \geq i \\ 0, & j < i \end{cases}$$

where the sum is the sum of the first  $n_L$  values and  $b_{ij}$  is the value on the last column. For the last row we obviously get sum 1.

We get the normalized PageRank  $\mathbf{R}^{(1)}$  by dividing  $\mathbf{R}^{(1)} = \mathbf{R}^{(2)} / \|\mathbf{R}^{(2)}\|_1$ .

□

□

## 4.2 Changes in a complete graph

Complete graphs or similar structures are common both as parts of a site and as a way between different sites to try and gain a better rank. An image of a complete graph with five nodes can be seen in Fig. 3. We recall that the system matrix for this system is:

$$\mathbf{A} = \frac{1}{4} \begin{bmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix}$$

Using this we get the inverse of this system as:

$$(\mathbf{I} - c\mathbf{A}^\top)^{-1} = \begin{bmatrix} \frac{3c-4}{c^2+3c-4} & \frac{-c}{c^2+3c-4} & \frac{-c}{c^2+3c-4} & \frac{-c}{c^2+3c-4} & \frac{-c}{c^2+3c-4} \\ \frac{-c}{c^2+3c-4} & \frac{3c-4}{c^2+3c-4} & \frac{-c}{c^2+3c-4} & \frac{-c}{c^2+3c-4} & \frac{-c}{c^2+3c-4} \\ \frac{-c}{c^2+3c-4} & \frac{-c}{c^2+3c-4} & \frac{3c-4}{c^2+3c-4} & \frac{-c}{c^2+3c-4} & \frac{-c}{c^2+3c-4} \\ \frac{-c}{c^2+3c-4} & \frac{-c}{c^2+3c-4} & \frac{-c}{c^2+3c-4} & \frac{3c-4}{c^2+3c-4} & \frac{-c}{c^2+3c-4} \\ \frac{-c}{c^2+3c-4} & \frac{-c}{c^2+3c-4} & \frac{-c}{c^2+3c-4} & \frac{-c}{c^2+3c-4} & \frac{3c-4}{c^2+3c-4} \end{bmatrix}$$

After normalization we will obviously end up with  $\mathbf{R}_i^{(1)} = 1/5$  as PageRank for every node  $i$ . However since there is not any dangling nodes in the complete graph all the nodes will have maximum influence on the PageRank of the system. Additionally since they only point to each other they will not share any of it with the outside in the case of a bigger link matrix with a part of it being a complete graph. This makes a complete graph similar to a dangling node in that it will not increase the rank of anyone else, but with the addition of having a higher rank in itself since it can increase its own rank to a certain extent.

Trying to find an expression for the elements in the inverse  $(\mathbf{I} - c\mathbf{A}^\top)^{-1}$  for the complete graph we formulate the following lemma:

**Lemma 4.2.** *The diagonal element  $a_d$  of the inverse  $(\mathbf{I} - c\mathbf{A}^\top)^{-1}$  of the complete graph with  $n$  nodes is:*

$$a_d = \frac{(n-1) - c(n-2)}{(n-1) - c(n-2) - c^2} \quad (24)$$

*The non diagonal elements  $a_{ij}$  can be written as:*

$$a_{ij} = \frac{c}{(n-1) - c(n-2) - c^2} \quad (25)$$

*Proof.* The diagonal element is the sum of the probabilities of all paths to node  $e_d$  from itself. This can be written as a geometric sum:  $a_d = \sum_{k=0}^{\infty} P(e_d \rightarrow e_d)^k$ , where  $P(e_d \rightarrow e_d)$

is the probability of getting from node  $e_d$  to node  $e_d$ . The probability  $P(e_d \rightarrow e_d)$  can be written as:

$$\begin{aligned} P(e_d \rightarrow e_d) &= \frac{c^2}{n-1} + \frac{c^3(n-2)}{(n-1)^2} + \frac{c^4(n-2)^2}{(n-1)^3} + \dots \\ &= \frac{c^2}{n-1} \sum_{k=0}^{\infty} \left( \frac{c(n-2)}{n-1} \right)^k = \frac{c^2}{(n-1) - c(n-2)} \end{aligned} \quad (26)$$

This gives:

$$a_d = \sum_{k=0}^{\infty} \left( \frac{c^2}{(n-1) - c(n-2)} \right)^k = \frac{(n-1) - c(n-2)}{(n-1) - c(n-2) - c^2} \quad (27)$$

For non-diagonal elements  $e_{ij}$  we get  $e_{ij} = P(e_i \rightarrow e_j)a_d$ , where  $P(e_i \rightarrow e_j)$  is the probability of getting from node  $e_i$  to node  $e_j$  where  $e_i \neq e_j$ . This probability can be written as:

$$\begin{aligned} P(e_i \rightarrow e_j) &= \frac{c}{n-1} + \frac{c^2(n-2)}{(n-1)^2} + \frac{c^3(n-2)^2}{(n-1)^3} + \dots \\ &= \frac{c}{n-1} \sum_{k=0}^{\infty} \left( \frac{c(n-2)}{n-1} \right)^k = \frac{c}{(n-1) - c(n-2)} \end{aligned} \quad (28)$$

This gives:

$$a_{ij} = \frac{c}{(n-1) - c(n-2)} \frac{(n-1) - c(n-2)}{(n-1) - c(n-2) - c^2} = \frac{c}{(n-1) - c(n-2) - c^2} \quad (29)$$

□

□

We give a matrix proof of Lemma 4.2 as well:

*Proof of Lemma 4.2.* We consider a general matrix  $\mathbf{A}$  of the form:

$$\mathbf{A} = \begin{bmatrix} 1 & a & a & \dots & a \\ a & 1 & a & \dots & a \\ a & a & 1 & \ddots & a \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ a & a & a & \dots & 1 \end{bmatrix}$$

We use Gauss-Jordan elimination to find the inverse  $\mathbf{A}^{-1}$ :

$$\begin{bmatrix} 1 & a & a & \dots & a & 1 & 0 & 0 & \dots & 0 \\ a & 1 & a & \dots & a & 0 & 1 & 0 & \dots & 0 \\ a & a & 1 & \ddots & a & 0 & 0 & 1 & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \vdots & \ddots & \ddots & 0 \\ a & a & a & \dots & 1 & 0 & 0 & 0 & \dots & 1 \end{bmatrix}$$



We add  $-ar_1$  where  $r_1$  is the first row to every other row to eliminate the elements below 1 on the first column.

$$\begin{bmatrix} 1 & a & a & \dots & a & 1 & 0 & 0 & \dots & 0 \\ 0 & 1-a^2 & a-a^2 & \dots & a-a^2 & -a & 1 & 0 & \dots & 0 \\ 0 & a-a^2 & 1-a^2 & \ddots & a-a^2 & -a & 0 & 1 & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & a-a^2 & a-a^2 & \dots & 1-a^2 & -a & 0 & 0 & \dots & 1 \end{bmatrix}$$

Next we eliminate the values to the right of the 1 on the first row. We add  $-k \sum_{i=2}^n r_i$ , where  $r_i$  is row  $i$  to the first row giving the equation:

$$\begin{aligned} a &= -k(1-a^2 + (n-2)(a-a^2)) \\ \Rightarrow k &= \frac{-a}{(1-a^2 + (n-2)(a-a^2))} \end{aligned} \quad (30)$$

This gives:

$$\begin{bmatrix} 1 & 0 & 0 & \dots & 0 & 1-(n-1)ak & k & k & \dots & k \\ 0 & 1-a^2 & a-a^2 & \dots & a-a^2 & -a & 1 & 0 & \dots & 0 \\ 0 & a-a^2 & 1-a^2 & \ddots & a-a^2 & -a & 0 & 1 & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & a-a^2 & a-a^2 & \dots & 1-a^2 & -a & 0 & 0 & \dots & 1 \end{bmatrix}$$

We are now done calculating the first row of the inverse  $A^{-1}$ . We get the other rows using the same calculations if we start with another pivot element. For the inverse matrix we get diagonal elements  $d = 1 - (n-1)ak$  and for all other elements  $e = k$ , where  $n$  is the total number of rows giving a inverse like below:

$$A^{-1} = \begin{bmatrix} 1-(n-1)ak & k & k & \dots & k \\ k & 1-(n-1)ak & k & \dots & k \\ k & k & 1-(n-1)ak & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & k \\ k & k & \dots & k & 1-(n-1)ak \end{bmatrix}$$

Calculating for  $a = -c/(n-1)$  as for a complete graph gives:

$$k = \frac{-a}{(1-a^2 + (n-2)(a-a^2))} = \frac{c}{(n-1) - (n-2)c - c^2} \quad (31)$$

$$\begin{aligned} d = 1 - (n-1)ak &= \frac{(n-1) - (n-2)c - c^2 - (n-1)(-c)/(n-1)c}{(n-1) - (n-2)c - c^2} \\ &= \frac{(n-1) - (n-2)c}{(n-1) - (n-2)c - c^2} \end{aligned} \quad (32)$$

And the proof is complete. □ □

Using this we immediately get the PageRank (before normalization) of elements in a complete graph with uniform  $\mathbf{u}$ :

**Theorem 4.3.** *Given a complete graph with  $n > 1$  nodes, PageRank  $\mathbf{R}^{(2)}$  before normalization can be written as:*

$$\mathbf{R}_i^{(2)} = \frac{1}{1-c} \quad (33)$$

*Proof.* From Lemma 4.2 We already have the inverse  $(\mathbf{I} - c\mathbf{A}^\top)^{-1}$ , We then find the PageRank by summation of any row of the matrix (since all rows have equal sum).

$$\begin{aligned} \mathbf{R}_i^{(2)} &= a_d + (n-1)a_{ij}, \quad i \neq j \\ &= \frac{(n-1) - c(n-2) + c(n-1)}{(n-1) - c(n-2) - c^2} = \frac{c + (n-1)}{(n-1) - c(n-2) - c^2} = \frac{1}{1-c} \end{aligned} \quad (34)$$

□

□

We do note that since we have no dangling nodes all the probability from a node in the complete graph is distributed within the complete graph. Also the size of the graph is irrelevant for the individual nodes as long as none are linked to from outside sources and it consists of at least two nodes. In the  $\mathbf{R}^{(1)}$  sense the size obviously changes the result since we would increase the overall number of nodes in the system by increasing the size of the complete graph. Two things is important to note however: The higher ones own PageRank before joining the complete graph (probability of getting there from outside nodes) the more gain there is by joining a small complete graph in order to maximize the probability of returning to itself. In the same way if a node have a very low rank it gains much by joining a large complete graph of nodes with higher rank than itself.

#### 4.2.1 Adding a link out of a complete graph

If we want to see how the complete graph changes when adding one link from one node (node one) out of the complete graph we end up with the following system matrix for the nodes in the complete graph:

$$(\mathbf{I} - c\mathbf{A}^\top) = \begin{bmatrix} 1 & -c/4 & -c/4 & -c/4 & -c/4 \\ -c/5 & 1 & -c/4 & -c/4 & -c/4 \\ -c/5 & -c/4 & 1 & -c/4 & -c/4 \\ -c/5 & -c/4 & -c/4 & 1 & -c/4 \\ -c/5 & -c/4 & -c/4 & -c/4 & 1 \end{bmatrix}$$

After taking the inverse and multiplying with  $-1$  we get:

$$(\mathbf{I} - c\mathbf{A}^\top)^{-1} =$$

$$\begin{bmatrix} \frac{15c-20}{-4c} & \frac{-5c}{12c^2+40c-80} & \frac{-5c}{4c(5+c)} & \frac{-5c}{4c(5+c)} & \frac{-5c}{4c(5+c)} \\ \frac{s}{-4c} & \frac{(c+4)s}{4c(5+c)} & \frac{(c+4)s}{12c^2+40c-80} & \frac{(c+4)s}{4c(5+c)} & \frac{(c+4)s}{4c(5+c)} \\ \frac{s}{-4c} & \frac{(c+4)s}{4c(5+c)} & \frac{(c+4)s}{4c(5+c)} & \frac{(c+4)s}{12c^2+40c-80} & \frac{(c+4)s}{4c(5+c)} \\ \frac{s}{-4c} & \frac{(c+4)s}{4c(5+c)} & \frac{(c+4)s}{4c(5+c)} & \frac{(c+4)s}{4c(5+c)} & \frac{(c+4)s}{12c^2+40c-80} \\ \frac{s}{-4c} & \frac{(c+4)s}{4c(5+c)} & \frac{(c+4)s}{4c(5+c)} & \frac{(c+4)s}{4c(5+c)} & \frac{(c+4)s}{12c^2+40c-80} \end{bmatrix}$$

where  $s = 4c^2 + 15c - 20$

We find the expression for the PageRank in a complete graph with one node linking out to be the following assuming uniform  $\mathbf{u}$ .

**Theorem 4.4.** *The PageRank of the nodes in a complete graph with the first node linking out of the complete graph, the PageRank can be written as:*

$$\mathbf{R}_1^{(2)} = \frac{n(n-1) + nc}{n(n-1) - (n-1)c^2 - n(n-2)c} \quad (35)$$

$$\mathbf{R}_i^{(2)} = \frac{(c+n)(n-1)}{n(n-1) - (n-1)c^2 - n(n-2)c}, \quad n \geq i > 1 \quad (36)$$

where  $n$  is the number of nodes in the complete graph and node one links out of the complete graph.

*Proof.* We start by looking at the PageRank as a probability, we let  $e_1$  be the node linking out. The probability to get from  $e_1$  back to itself is:

$$\begin{aligned} P(e_1 \rightarrow e_1) &= \frac{c(n-1)}{n} \frac{c}{n-1} + \frac{c(n-1)}{n} \frac{c}{n-1} \frac{c(n-2)}{n-1} \\ &\quad + \frac{c(n-1)}{n} \frac{c}{n-1} \left( \frac{c(n-2)}{n-1} \right)^2 + \dots \\ &= \frac{c^2}{n} \sum_{k=0}^{\infty} \left( \frac{c(n-2)}{n-1} \right)^k = \frac{c^2}{n} \frac{n-1}{(n-1) - c(n-2)} \end{aligned} \quad (37)$$

And we get the sum of all probabilities from  $e_1$  back to itself as:

$$\begin{aligned} \sum_{k=0}^{\infty} (P(e_1 \rightarrow e_1))^k &= \sum_{k=0}^{\infty} \left( \frac{c^2}{n} \frac{n-1}{(n-1) - c(n-2)} \right)^k \\ &= \frac{n((n-1) - c(n-2))}{n((n-1) - c(n-2)) - c^2(n-1)} = \mathbf{B}^{inv} \end{aligned} \quad (38)$$

We remember that on the diagonal of  $(\mathbf{I} - c\mathbf{A}^\top)$ , we have the sums of probabilities of nodes going back to themselves. So if we divide the matrix  $(\mathbf{I} - c\mathbf{A}^\top)$  in blocks:

$$(\mathbf{I} - c\mathbf{A}^\top) = \begin{bmatrix} B & C \\ D & E \end{bmatrix}$$

And inverse matrix:

$$(\mathbf{I} - c\mathbf{A}^\top)^{-1} = \begin{bmatrix} \mathbf{B}^{inv} & \mathbf{C}^{inv} \\ \mathbf{D}^{inv} & \mathbf{E}^{inv} \end{bmatrix}$$

We note that  $\mathbf{B}^{inv}$  is not the inverse of  $\mathbf{B}$  but the part of the inverse  $(\mathbf{I} - c\mathbf{A}^\top)^{-1}$  corresponding to block  $\mathbf{B}$ . We let  $\mathbf{B} = [1]$  corresponding to the node linking out and we get  $\mathbf{B}^{inv}$  as above.

For the elements  $C_i^{inv}, i \neq 1$  of  $\mathbf{C}^{inv}$  we find them as

$$\begin{aligned} C_i^{inv} &= \sum_{k=0}^{\infty} (P(e_i \rightarrow e_1))^k \sum_{k=0}^{\infty} (P(e_1 \rightarrow e_1))^k \\ &= \frac{c}{n-1} \sum_{k=0}^{\infty} \left( \frac{c(n-2)}{n-1} \right)^k \mathbf{B}^{inv} = \frac{cn}{n((n-1) - c(n-2)) - c^2(n-1)} \end{aligned} \quad (39)$$

Since  $\mathbf{E}$  and  $\mathbf{DB}^{-1}\mathbf{C}$  are both symmetric and have every non-diagonal element equal as well as all diagonal elements equal, the inverse  $\mathbf{E}^{inv} = (\mathbf{E} - \mathbf{DB}^{-1}\mathbf{C})^{-1}$  should be the same as well. Especially every row and column have the same sum. From Lemma 4.1 for blockwise inversion we get:

$$C_i^{inv} = -\frac{-c}{n-1} \sum_{k=1}^{n-1} E_{ki}^{inv} \quad (40)$$

$$D_i^{inv} = -\frac{-c}{n} \sum_{k=1}^{n-1} E_{ik}^{inv} = -\frac{-c}{n} \sum_{k=1}^{n-1} E_{ki}^{inv} \quad (41)$$

$$\Rightarrow \begin{cases} D_i^{inv} = \frac{(n-1)C_i^{inv}}{n} \\ \sum_{k=1}^{n-1} E_{ik}^{inv} = \frac{(n-1)C_i^{inv}}{c} \end{cases} \quad (42)$$

We get the PageRank as:

$$\begin{aligned} \mathbf{R}_1^{(2)} &= \mathbf{B}^{inv} + (n-1)\mathbf{C}^{inv} = \frac{n((n-1) - c(n-2))}{n((n-1) - c(n-2)) - c^2(n-1)} \\ &+ \frac{(n-1)cn}{n((n-1) - c(n-2)) - c^2(n-1)} = \frac{n(n-1) + nc}{n(n-1) - (n-1)c^2 - n(n-2)c} \end{aligned} \quad (43)$$

$$\begin{aligned} \mathbf{R}_i^{(2)} &= D^{inv} + \sum_{k=1}^{n-1} E_{ik}^{inv} = \frac{(n-1)C_i^{inv}}{n} + \frac{(n-1)C_i^{inv}}{c} \\ &= \frac{(n-1)C_i^{inv}(c+n)}{nc} = \frac{(c+n)(n-1)}{n(n-1) - (n-1)c^2 - n(n-2)c} \end{aligned} \quad (44)$$

And the proof is complete.  $\square$   $\square$

We give a matrix proof of Theorem 4.4 as well:

*Proof of Theorem 4.4.* We consider the square matrix  $A$  with  $n$  rows.

$$A = \begin{bmatrix} 1 & a & a & \dots & a \\ b & 1 & a & \dots & a \\ b & a & 1 & \ddots & a \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ b & a & a & \dots & a \end{bmatrix}$$

Where  $a = -c/(n-1)$ ,  $b = -c/n$ . We divide the matrix in blocks:

$$A = \begin{bmatrix} B & C \\ D & E \end{bmatrix}$$

Where  $B = [1]$ ,  $C = [a \ a \ \dots \ a]$ ,  $D = [b \ b \ \dots \ b]^\top$  and  $E$  have looks like the matrix for a complete graph but is of size  $(n-1) \times (n-1)$ :

$$E = \begin{bmatrix} 1 & a & a & \dots & a \\ a & 1 & a & \dots & a \\ a & a & 1 & \ddots & a \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ a & a & a & \dots & a \end{bmatrix}$$

In the same way as in the proof of Lemma 4.2 we find the elements of  $B, C$  by choosing the top left element as pivot element. This gives

$$k_A = \frac{-a}{(1-ab) + (n-2)(a-ab)} \quad (45)$$

We write  $A^{-1}$  as blocks:

$$A^{-1} = \begin{bmatrix} B^{inv} & C^{inv} \\ D^{inv} & E^{inv} \end{bmatrix}$$

and get:  $B^{inv} = 1 - (n-1)bk_A$  and  $C_i^{inv} = k_A$ .

From the matrix proof of Lemma 4.2 we get the non-diagonal elements  $E_e$  and diagonal elements  $E_d$  of  $E^{-1}$  as

$$E_e = k_D = \frac{-a}{(1-a^2) + (n-3)(a-a^2)} = \frac{(n-1)c}{(n-1)^2 - (n-3)(n-1)c - (n-2)c^2} \quad (46)$$

$$E_d = 1 - (n-2)ak_D = \frac{(n-1)^2 - (n-3)(n-1)c}{(n-1)^2 - (n-3)(n-1)c - (n-2)c^2} \quad (47)$$

From Lemma 4.1 we then get:

$$B^{inv} = (B - CE^{-1}D)^{-1} = 1 - (n-1)bk_A \quad (48)$$

$$\begin{aligned} C^{inv} &= -(B - CE^{-1}D)^{-1}CE^{-1} \\ \Rightarrow C_i^{inv} &= -(B - CE^{-1}D)^{-1}b(E_d + (n-2)E_e) = k_A \end{aligned} \quad (49)$$

$$\begin{aligned} D^{inv} &= -E^{-1}D(B - CE^{-1}D)^{-1} \\ \Rightarrow D_i^{inv} &= -a(E_d + (n-2)E_e)(B - CE^{-1}D)^{-1} = \frac{bk_A}{a} \end{aligned} \quad (50)$$

$$E^{inv} = E^{-1} + E^{-1}D(B - CE^{-1}D)^{-1}CE^{-1} \quad (51)$$

$$\Rightarrow \begin{cases} E_d^{inv} = E_d + b(E_d + (n-2)E_e)(B - CE^{-1}D)^{-1}a(E_d + (n-2)E_e) \\ \quad = E_d - b(E_d + (n-2)E_e)k_A \\ E_e^{inv} = E_e + b(E_d + (n-2)E_e)(B - CE^{-1}D)^{-1}a(E_d + (n-2)E_e) \\ \quad = E_e - b(E_d + (n-2)E_e)k_A \end{cases} \quad (52)$$

We replace  $a = -c/(n-1)$  and  $b = -c/n$  as for our complete graph and get inverse:

$$\begin{aligned} & (I - cA^\top)^{-1} \\ &= \begin{bmatrix} 1 - (n-1)bk_A & k_A & k_A & \dots & k_A \\ \frac{bk_A}{a} & 1 - (n-2)ak_D & k_D & \dots & k_D \\ \frac{bk_A}{a} & k_D & 1 - (n-2)ak_D & \ddots & k_D \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \frac{bk_A}{a} & k_D & k_D & \dots & 1 - (n-2)ak_D \end{bmatrix} \end{aligned}$$

For the PageRank of the node linking out we get:

$$\begin{aligned} \mathbf{R}_1^{(2)} &= \mathbf{B}^{inv} + (n-1)C_i^{inv} = 1 - (n-1)bk_A + (n-1)k_A = 1 - (n-1)(b-1)k_A \\ &= \frac{(1-ab) + (n-2)(a-ab) + (n-1)(b-1)a}{(1-ab) + (n-2)(a-ab)} \\ &= \frac{(1-ab) - (a-ab) - (n-1)a}{(1-ab) + (n-2)(a-ab)} = \frac{n(n-1) + cn}{n(n-1) - n(n-2)c - (n-1)c^2} \end{aligned} \quad (53)$$

For all other nodes we get PageRank:

$$\begin{aligned} \mathbf{R}_i^{(2)} &= D_i^{inv} + E_d^{inv} + (n-2)E_e^{inv} = \frac{bk_A}{a} + E_d - b(E_d + (n-2)E_e)k_A \\ &\quad + (n-2)E_e - (n-2)b(E_d + (n-2)E_e)k_A \\ &= E_d + (n-2)E_e - (n-1)b(E_d + (n-2)E_e)k_A + (b/a)k_A \\ &= \frac{1-a}{(1-a^2) + (n-3)(a-a^2)} + \frac{-b}{(1-ab) + (n-2)(a-ab)} \\ &\quad + \frac{(n-1)ab(1-a)}{((1-a^2) + (n-3)(a-a^2))((1-ab) + (n-2)(a-ab))} \\ &= \frac{1-b}{1-ab + (n-2)(a-ab)} = \frac{(n-1)(n+c)}{n(n-1) - n(n-2)c - (n-1)c^2} \end{aligned} \quad (54)$$

And the proof is complete. □

Just looking at the expression it is hard to say how the PageRank changes after linking out. We can however note a couple of things: First of all the PageRank is lower than for the complete graph (since we now have a chance to escape the graph). But more interesting, when comparing the node that links out with the others we formulate the following theorem:

**Theorem 4.5.** *In a complete graph not linked to from the outside but with one link out, the node that links out will have the highest PageRank in the complete graph.*

*Proof.* Using the expression for PageRank in a complete graph with one link out we want to prove  $\mathbf{R}_k^{(2)} > \mathbf{R}_i^{(2)}$  where  $\mathbf{R}_k^{(2)}$  is the PageRank for the node linking out and  $\mathbf{R}_i^{(2)}$  is the PageRank of all the other nodes.

$$\begin{aligned}
& \mathbf{R}_k^{(2)} > \mathbf{R}_i^{(2)} \\
\Leftrightarrow & \frac{n(n-1) + nc}{n(n-1) - (n-1)c^2 - n(n-2)c} > \frac{(c+n)(n-1)}{n(n-1) - (n-1)c^2 - n(n-2)c} \quad (55) \\
& \Leftrightarrow n(n-1) + nc > (c+n)(n-1) \\
& \Leftrightarrow n^2 + nc - n > n^2 + nc - n - c
\end{aligned}$$

Where  $0 < c < 1$  and  $n > 1$  is the number of nodes in the complete graph. This is obviously true and the proof is complete. □

Apart from the knowledge that it is the node that links out of a complete graph that loses the least from it we can also see that as the number of nodes in the complete graph increases the difference between them decreases since we have a factor  $n^2$  in the denominator compared to only a difference of  $c$  in the nominator.

#### 4.2.2 Effects of linking to a complete graph

In the case of a link to a complete graph without a link back from the complete graph we can easily guess the result. From earlier we know that for a node linking to one other node in a link matrix with no change of getting back to itself the column corresponding to the node linking out is  $c$  times the column of the node it links to. Additionally we need to add a one to the diagonal element for that column.

The fact that there is no probability (or a very low if it is only close to complete) to escape the complete graph and give any advantage to other parts of the system means the complete graph as a whole get maximum benefit from the links to it. Looking at how the additional probability  $c/(1-c) = c + c^2 + c^3 + \dots + c^\infty$  get distributed within the complete graph we realize that the node linked to gains all of the initial  $c^1$ , then loses a part  $c^2$  distributed among all other nodes in the complete graph, after that the rest is close to evenly distributed between all the nodes in the complete graph. As such we see that the node linked to is the node which gains the most from the link (which is what we would expect).

### 4.3 Connecting the simple line with the complete graph

Here we will look at what happens when we connect a complete graph with a simple line in various ways. This way we can get some information on what type of structure is most effective in getting a high PageRank and see how they interact with each other.

#### 4.3.1 Connecting the simple line with a link from a node in the complete graph to a node in the line

Looking at the system where we let one node in a complete graph link to one node in a simple line we get a system similar to the case where we added a single node to the line (complete graph with one node). An example of what the system could look like can be seen in Fig. 5. We have the two systems  $S_L$ ,  $S_G$  as the original systems for the simple line and complete graph respectively. We want to find the new PageRank of these nodes after creating our new system  $S$  by adding a link from the first node in the complete graph  $e_{G,1}$  to node  $e_{L,j}$  in the simple line. When using  $n_L = 5$ ,  $n_G = 5$ ,  $j = 3$  we get the system with  $(I - cA^\top)$  seen below:

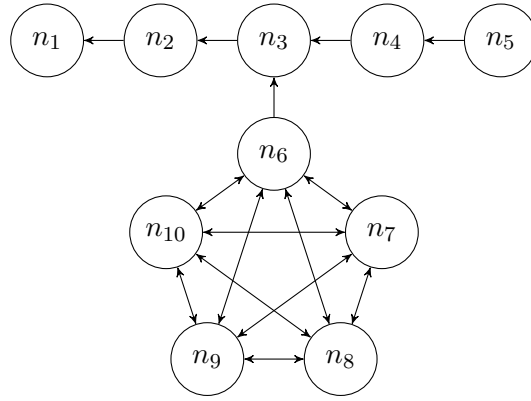


Figure 5: Simple line with one link from a complete graph to one node in the line

$$I - cA^\top = \begin{bmatrix} 1 & -c & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -c & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -c & 0 & -c/5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -c & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -c/4 & -c/4 & -c/4 & -c/4 \\ 0 & 0 & 0 & 0 & 0 & -c/5 & 1 & -c/4 & -c/4 & -c/4 \\ 0 & 0 & 0 & 0 & 0 & -c/5 & -c/4 & 1 & -c/4 & -c/4 \\ 0 & 0 & 0 & 0 & 0 & -c/5 & -c/4 & -c/4 & 1 & -c/4 \\ 0 & 0 & 0 & 0 & 0 & -c/5 & -c/4 & -c/4 & -c/4 & 1 \end{bmatrix}$$

We find the inverse as:

$$(I - cA^\top)^{-1} =$$



$$\begin{bmatrix} 1 & c & c^2 & c^3 & c^4 & \frac{c^3(3c-4)}{s} & -\frac{c^4}{t} & -\frac{c^4}{t} & -\frac{c^4}{t} & -\frac{c^4}{t} \\ 0 & 1 & c & c^2 & c^3 & \frac{c^2(3c-4)}{s} & -\frac{c^3}{t} & -\frac{c^3}{t} & -\frac{c^3}{t} & -\frac{c^3}{t} \\ 0 & 0 & 1 & c & c^2 & \frac{c(3c-4)}{s} & -\frac{c^2}{t} & -\frac{c^2}{t} & -\frac{c^2}{t} & -\frac{c^2}{t} \\ 0 & 0 & 0 & 1 & c & \frac{s}{0} & \frac{s}{0} & \frac{s}{0} & \frac{s}{0} & \frac{s}{0} \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{15c-20}{s} & -\frac{5c}{t} & -\frac{5c}{t} & -\frac{5c}{t} & -\frac{5c}{t} \\ 0 & 0 & 0 & 0 & 0 & -\frac{4c}{s} & \frac{s}{t} & -\frac{4c(\frac{s}{5}+c)}{t} & -\frac{4c(\frac{s}{5}+c)}{t} & -\frac{4c(\frac{s}{5}+c)}{t} \\ 0 & 0 & 0 & 0 & 0 & \frac{s}{-4c} & \frac{(c+4)s}{4c(5+c)} & -\frac{(c+4)s}{t} & -\frac{(c+4)s}{4c(5+c)} & -\frac{(c+4)s}{4c(5+c)} \\ 0 & 0 & 0 & 0 & 0 & -\frac{4c}{s} & -\frac{(c+4)s}{4c(5+c)} & \frac{(c+4)s}{4c(5+c)} & -\frac{(c+4)s}{t} & -\frac{(c+4)s}{4c(5+c)} \\ 0 & 0 & 0 & 0 & 0 & \frac{s}{-4c} & -\frac{(c+4)s}{4c(5+c)} & -\frac{(c+4)s}{4c(5+c)} & \frac{(c+4)s}{t} & -\frac{(c+4)s}{4c(5+c)} \\ 0 & 0 & 0 & 0 & 0 & -\frac{4c}{s} & \frac{(c+4)s}{4c(5+c)} & -\frac{(c+4)s}{4c(5+c)} & -\frac{(c+4)s}{4c(5+c)} & \frac{(c+4)s}{t} \end{bmatrix}$$

where  $s = 4c^2 + 15c - 20$  and  $t = 12c^2 + 40c - 20$ . Using what we earlier learned by doing changes to the simple line and complete graph separately we can note a couple of things. For the nodes in the complete graph we get the same expression as when adding a link out of the complete graph (since there are no link back to the complete graph).

Assuming uniform  $\mathbf{u}$  the PageRank in the simple line after adding the link from the complete graph  $\mathbf{R}_L^{(2)}[S_G \rightarrow S_L]$  can still be written in about the same way:

**Theorem 4.6.** *Observing the nodes in a system  $S$  made up of two systems, a simple line  $S_L$  with  $n_L$  nodes and a complete graph  $S_G$  with  $n_G$  nodes where we add one link from node  $e_g$  in the complete graph to node  $e_j$  in the simple line. Assuming uniform  $\mathbf{u}$  we get the PageRank  $\mathbf{R}_{L,i}^{(2)}[S_G \rightarrow S_L]$  for the nodes in the line after the new link and  $\mathbf{R}_{G,i}^{(2)}[S_G \rightarrow S_L]$  for the nodes in the complete graph after the new link as:*

$$\begin{aligned} \mathbf{R}_{L,i}^{(2)}[S_G \rightarrow S_L] &= \sum_{k=0}^{n_L-i} c^k + b_{ij} = \frac{1 - c^{n_L-i+1}}{1 - c} + b_{ij} \\ b_{ij} &= -c^{j+1-i} \frac{c + (n_G - 1)}{(n_G - 1)c^2 + n_G(n_G - 2)c - n_G(n_G - 1)}, \quad j \geq i \\ b_{ij} &= 0, \quad j < i \end{aligned} \tag{56}$$

For the nodes in the complete graph we get:

$$\mathbf{R}_{G,1}^{(2)}[S_G \rightarrow S_L] = -\frac{n_G(n_G - 1) + n_G c}{(n_G - 1)c^2 + n_G(n_G - 2)c - n_G(n_G - 1)} \tag{57}$$

$$\mathbf{R}_{G,j}^{(2)}[S_G \rightarrow S_L] = \frac{(c + n_G)(n_G - 1)}{n_G(n_G - 1) - (n_G - 1)c^2 - n_G(n_G - 2)c} \tag{58}$$

where  $\mathbf{R}_{G,1}^{(2)}[S_G \rightarrow S_L]$  is PageRank for the node in the complete graph linking to the line and  $\mathbf{R}_{G,j}^{(2)}[S_G \rightarrow S_L]$  is the PageRank of the other nodes in the complete graph.

*Proof.* For the nodes in the complete graph we get the PageRank immediately from Theorem 4.4.

For the nodes in the line we get a similar result as when adding a link from a single node to the line in Theorem 4.2. We get the same PageRank for the nodes we can not reach from the complete graph ( $b_{ij} = 0$ ,  $j < i$ ). For the nodes we can reach we need to modify  $b_{ij}$ . The sum of all probability to reach the node in the complete graph linking to the line is found in equation 35 in Theorem 4.4.

$$\mathbf{R}_{G,1}^{(2)}[S_G \rightarrow S_L] = -\frac{n_G(n_G - 1) + n_G c}{(n_G - 1)c^2 + n_G(n_G - 2)c - n_G(n_G - 1)}$$

The probability to reach the linked to node in the line  $e_j$  is then

$$\left(\frac{c}{n_G}\right) \mathbf{R}_{e_1 \in S_G}^{(2)}[S_G \rightarrow S_L]$$

and for any further node in the line we need to multiply with  $c$  for every extra step. This gives:

$$\begin{aligned} b_{ij} &= -c^{j-i} \frac{c}{n_G} \frac{n_G(n_G - 1) + n_G c}{(n_G - 1)c^2 + n_G(n_G - 2)c - n_G(n_G - 1)} \\ &= -c^{j+1-i} \frac{c + (n_G - 1)}{(n_G - 1)c^2 + n_G(n_G - 2)c - n_G(n_G - 1)}, \quad j \geq i \end{aligned} \quad (59)$$

And the proof is complete.  $\square$   $\square$

We give a matrix proof as well:

*Proof of Theorem 4.6.* We divide  $(\mathbf{I} - c\mathbf{A}^\top)$  in blocks:

$$(\mathbf{I} - c\mathbf{A}^\top) = \begin{bmatrix} \mathbf{B} & \mathbf{C} \\ \mathbf{D} & \mathbf{E} \end{bmatrix}$$

where  $\mathbf{B}$  is a  $n_L \times n_L$  matrix corresponding to the nodes in the line,  $\mathbf{C}$  is a  $n_G \times n_L$  matrix of all elements zero except element  $C_{jg} = -c/n_G$ , where  $e_j$  is the node in the line linked to by  $e_g$  in the complete graph.  $\mathbf{D}$  is a zero matrix of size  $n_L \times n_G$  and  $\mathbf{E}$  is the  $n_G \times n_G$  matrix corresponding to a complete graph with node  $e_g$  linking out of the graph. We write:

$$(\mathbf{I} - c\mathbf{A}^\top)^{-1} = \begin{bmatrix} \mathbf{B}^{inv} & \mathbf{C}^{inv} \\ \mathbf{D}^{inv} & \mathbf{E}^{inv} \end{bmatrix}$$

From Lemma 4.1 for blockwise inversion we get:

$$\begin{aligned} \mathbf{B}^{inv} &= (\mathbf{B} - \mathbf{D}\mathbf{E}^{-1}\mathbf{C})^{-1} = \mathbf{B}^{-1} \\ \mathbf{C}^{inv} &= -(\mathbf{B} - \mathbf{D}\mathbf{E}^{-1}\mathbf{C})^{-1}\mathbf{C}\mathbf{E}^{-1} = -\mathbf{B}^{-1}\mathbf{C}\mathbf{E}^{-1} \\ \mathbf{D}^{inv} &= -\mathbf{E}^{-1}\mathbf{D}(\mathbf{B} - \mathbf{D}\mathbf{E}^{-1}\mathbf{C})^{-1} = 0 \\ \mathbf{E}^{inv} &= \mathbf{E}^{-1} + \mathbf{E}^{-1}\mathbf{D}(\mathbf{B} - \mathbf{D}\mathbf{E}^{-1}\mathbf{C})^{-1}\mathbf{C}\mathbf{E}^{-1} = \mathbf{E}^{-1} \end{aligned} \quad (60)$$

Since  $D = 0$  and  $E$  is the matrix for a complete graph with a node linking out we get from Theorem 4.4 the PageRank for the nodes in the complete graph:

$$\mathbf{R}_{G,g}^{(2)}[S_G \rightarrow S_L] = \frac{n(n-1) + nc}{n(n-1) - (n-1)c^2 - n(n-2)c}$$

$$\mathbf{R}_{G,i}^{(2)}[S_G \rightarrow S_L] = \frac{(c+n)(n-1)}{n(n-1) - (n-1)c^2 - n(n-2)c}, \quad i \neq g$$

For the nodes in the line we need to calculate  $C^{inv}$ .

$$C^{inv} = -B^{-1}CE^{-1} = - \begin{bmatrix} 1 & c & c^2 & \dots & c^{n_L-1} \\ 0 & 1 & c & \dots & c^{n_L-2} \\ 0 & 0 & 1 & \ddots & c^{n_L-3} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix} CE^{-1}$$

Calculating  $-B^{-1}C$  we get:

$$(-B^{-1}C)_{kl} = \begin{cases} 0, & l \neq g \\ 0, & k > j \\ c^{j-k}c/n_G, & k \leq j, \quad l = g \end{cases} \quad (61)$$

In other words zero except for column  $g$ . Since only one column is non-zero the sum of every row of  $-B^{-1}CE^{-1}$  is then easily found as:

$$\sum_{l=1}^{n_G} (-B^{-1}CE^{-1})_{kl} = \begin{cases} \frac{c^{j-k+1}}{n_G} \sum_{l=1}^{n_G} E_{kl}^{-1} = \frac{c^{j-k+1}}{n_G} \mathbf{R}_{G,g}^{(2)}[S_G \rightarrow S_L], & k \leq j \\ 0, & k > j \end{cases} \quad (62)$$

Since  $B$  is the matrix for a line we get the total PageRank of the nodes in the line as:

$$\mathbf{R}_{L,i}^{(2)}[S_G \rightarrow S_L] = \frac{1 - c^{n_L-i+1}}{1 - c} + \frac{c^{j-k}c}{n_G} \sum_{l=1}^{n_G} E_{kl}^{-1} \quad (63)$$

Where the first part is the part corresponding to the PageRank of a line and the second part is the part influenced by the complete graph. Calculating  $\frac{c^{j-k}c}{n_G} \sum_{l=1}^{n_G} E_{kl}^{-1}$  we get:

$$\frac{c^{j-k}c}{n_G} \sum_{l=1}^{n_G} E_{kl}^{-1} = \begin{cases} \frac{-c^{j+1-i}(c + (n_G - 1))}{(n_G - 1)c^2 + n_G(n_G - 2)c - n_G(n_G - 1)}, & j \geq i \\ 0, & j < i \end{cases} \quad (64)$$

We replace  $\frac{c^{j-k}c}{n_G} \sum_{l=1}^{n_G} E_{kl}^{-1}$  with  $b_{ij}$  and the proof is complete.

For reference we include the whole inverse matrix as well (assuming the first node in the complete graph links out):

$$(I - cA^\top)^{-1} = \begin{bmatrix} \mathbf{B}^{\text{inv}} & \mathbf{C}^{\text{inv}} \\ \mathbf{D}^{\text{inv}} & \mathbf{E}^{\text{inv}} \end{bmatrix}, \quad \mathbf{B}^{\text{inv}} = \begin{bmatrix} 1 & c & c^2 & \dots & c^{n_L-1} \\ 0 & 1 & c & \dots & c^{n_L-2} \\ 0 & 0 & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & c \\ 0 & 0 & \dots & 0 & 1 \end{bmatrix}$$

$$\mathbf{C}^{\text{inv}} = \begin{bmatrix} \frac{c^j(1 - (n_G - 1)bk_A)}{n_G} & \frac{c^j k_A}{c^{j-1}k_A} & \frac{c^j k_A}{c^{j-1}k_A} & \dots & \frac{c^j k_A}{c^{j-1}k_A} \\ \frac{c^{j-1}(1 - (n_G - 1)bk_A)}{n_G} & \frac{c^{j-1}k_A}{c^{j-2}k_A} & \frac{c^{j-1}k_A}{c^{j-2}k_A} & \dots & \frac{c^{j-1}k_A}{c^{j-2}k_A} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ \frac{c(1 - (n_G - 1)bk_A)}{n_G} & \frac{ck_A}{n_G} & \frac{ck_A}{n_G} & \dots & \frac{ck_A}{n_G} \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}$$

$$\mathbf{D}^{\text{inv}} = \begin{bmatrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{bmatrix}, \quad \mathbf{E}^{\text{inv}} = \begin{bmatrix} \frac{E_A}{bk_A} & k_A & k_A & \dots & k_A \\ \frac{bk_A}{a} & E_D & k_D & \dots & k_D \\ \frac{bk_A}{a} & k_D & E_D & \ddots & k_D \\ \frac{bk_A}{a} & \vdots & \ddots & \ddots & \vdots \\ \frac{bk_A}{a} & k_D & k_D & \dots & E_D \end{bmatrix}$$

$$E_D = 1 - (n_G - 2)ak_D, \quad E_A = 1 - (n_G - 1)bk_A$$

$$k_A = \frac{-a}{(1 - ab) + (n_G - 2)(a - ab)}, \quad k_D = \frac{-a}{(1 - a^2) + (n_G - 3)(a - a^2)}$$

$$a = \frac{-c}{n_G - 1}, \quad b = \frac{-c}{n_G}$$

□

□

If we want to know the common normalized PageRank we find the normalizing con-

stant as the sum of the PageRank of all the nodes:

$$\begin{aligned}
N = \frac{n_L}{1-c} - \frac{c(1-c^{n_L})}{(1-c)^2} + \frac{c(1-c^{n_L-i+2})(c+n_G-1)}{(1-c)((n_G-1)c^2 + n_G(n_G-2)c - n_G(n_G-1))} \\
+ n_G(n_G-1) + n_G c(n_G-1)c^2 + n_G(n_G-2)c - n_G(n_G-1) \\
+ n_G \left( (n_G-1)c^2 + (2n_G(n_G-1)-1)c + (n_G(n_G-1))^2 \right) \\
\left( c \left( (n_G-1)c^2 + n_G(n_G-2)c - n_G(n_G-1) \right) \right. \\
\left. + n_G \left( (n_G-1)c^2 + n_G(n_G-2)c - n_G(n_G-1) \right) - 1 \right)^{-1}
\end{aligned} \tag{65}$$

Which can be used to get the normalized PageRank:

$$\mathbf{R}_i^{(1)}[S_G \rightarrow S_L] = \mathbf{R}_i^{(2)}[S_G \rightarrow S_L]/N \tag{66}$$

#### 4.3.2 Connecting the simple line with a complete graph by adding a link from a node in the line to a node in the complete graph

When we instead let one node  $e_j$  in the simple line link to one node in the complete graph we get a system that could look like the system in Fig. 6.

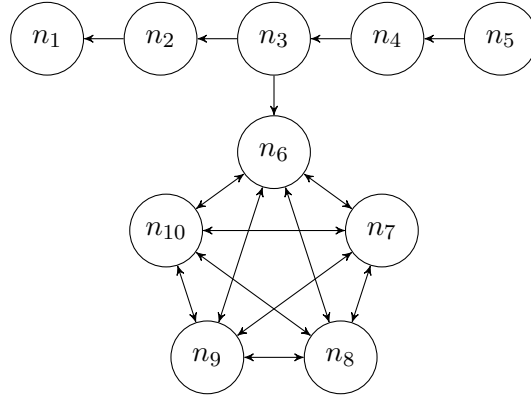


Figure 6: Simple line with one node in the line linking to a node in a complete graph

For the PageRank we formulate the following:

**Theorem 4.7.** *Observing the nodes in a system  $S$  made up of two systems, a simple line  $S_L$  with  $n_L$  nodes and a complete graph  $S_G$  with  $n_G$  nodes where we add one link from node  $e_j$  in the line to node  $e_g$  in the complete graph. Assuming uniform  $\mathbf{u}$  we get the PageRank  $\mathbf{R}_{L,i}^{(2)}[S_L \rightarrow S_G]$  for the nodes in the line after the new link and  $\mathbf{R}_{G,i}^{(2)}[S_L \rightarrow S_G]$  for the nodes in the complete graph after the new link as:*

$$\mathbf{R}_{L,i}^{(2)}[S_L \rightarrow S_G] = \frac{1 - c^{n_L+1-i}}{1 - c}, \quad i \geq j \quad (67)$$

$$\mathbf{R}_{G,g}^{(2)}[S_L \rightarrow S_G] = \left( \frac{c(1 - c^{n_L+1-j})}{2(1 - c)} \right) \left( \frac{((n_G - 1) - c(n_G - 2))}{((n_G - 1) - c(n_G - 2)) - c^2} \right) + \frac{1}{1 - c} \quad (68)$$

$$\mathbf{R}_{G,i}^{(2)}[S_L \rightarrow S_G] = \left( \frac{c^2(1 - c^{n_L+1-j})}{2(1 - c)} \right) \left( \frac{1}{((n_G - 1) - c(n_G - 2)) - c^2} \right) + \frac{1}{1 - c} \quad (69)$$

$$\mathbf{R}_{L,i}^{(2)}[S_L \rightarrow S_G] = \frac{1 - c^{j-i}}{1 - c} + \left( \frac{c^{j-i}}{2} \right) \frac{1 - c^{n_L-j+1}}{1 - c}, \quad i < j \quad (70)$$

*Proof.* For the nodes above the line we get the same PageRank as for the nodes "above" the linked to node in the line in Theorem 4.2:

$$\mathbf{R}_{L,i}^{(2)}[S_L \rightarrow S_G] = \frac{1 - c^{n_L+1-i}}{1 - c}, \quad i \geq j \quad (71)$$

In order to find  $\mathbf{R}_{G,g}^{(2)}[S_L \rightarrow S_G]$  we first write it as:

$$\mathbf{R}_{G,g}^{(2)}[S_L \rightarrow S_G] = \left( \sum_{e_i \in S, e_i \neq e_g} P(e_i \rightarrow e_g) + 1 \right) \left( \sum_{k=0}^{\infty} (P(e_g \rightarrow e_g))^k \right) \quad (72)$$

where  $P(e_i \rightarrow e_g)$  is the probability of getting from node  $e_i$  to node  $e_g$ .

$$\begin{aligned} \sum_{e_i \in S, e_i \neq e_g} P(e_i \rightarrow e_g) + 1 &= 1 + \frac{c}{2} \sum_{k=0}^{n_L-j} c^k + (n_G - 1) \frac{c}{n_G - 1} \sum_{k=0}^{\infty} \left( \frac{c(n_G - 2)}{(n_G - 1)} \right)^k \\ &= 1 + \frac{c(1 - c^{n_L+1-j})}{2(1 - c)} + \frac{c(n_G - 1)}{(n_G - 1) - c(n_G - 2)} \end{aligned} \quad (73)$$

$$P(e_g \rightarrow e_g) = \frac{c^2}{n_G - 1} \sum_{k=0}^{\infty} \left( \frac{c(n_G - 2)}{n_G - 1} \right)^k = \frac{c^2}{(n_G - 1) - c(n_G - 2)} \quad (74)$$

$$\sum_{k=0}^{\infty} (P(e_g \rightarrow e_g))^k = \frac{((n_G - 1) - c(n_G - 2))}{((n_G - 1) - c(n_G - 2)) - c^2} \quad (75)$$

Multiply the two expressions and we get the PageRank for the node linked to in the complete graph.

$$\begin{aligned} \mathbf{R}_{G,g}^{(2)}[S_L \rightarrow S_G] &= \left( \frac{c(1 - c^{n_L+1-j})}{2(1 - c)} + 1 + \frac{c(n_G - 1)}{(n_G - 1) - c(n_G - 2)} \right) \left( \frac{((n_G - 1) - c(n_G - 2))}{((n_G - 1) - c(n_G - 2)) - c^2} \right) \\ &= \left( \frac{c(1 - c^{n_L+1-j})}{2(1 - c)} \right) \left( \frac{((n_G - 1) - c(n_G - 2))}{((n_G - 1) - c(n_G - 2)) - c^2} \right) + \frac{1}{1 - c} \end{aligned} \quad (76)$$

We previously found the part equal to  $\frac{1}{1-c}$  after the proofs of Lemma 4.2.

Below  $e_j$  in the line we get a sum of two lines, half the probability of the line starting in  $e_L$  and all from the one starting in  $e_{j-1}$ , since node  $e_j$  have two links out where only one links to the line.

Collecting the sum of the probability of all nodes from  $e_j$  and above in one term and all below in one gives:

$$\mathbf{R}_{L,i}^{(2)}[S_L \rightarrow S_G] = \sum_{k=0}^{j-i-1} c^k + \sum_{k=j-i}^{n_L-i} \frac{c^k}{2} = \frac{1-c^{j-i}}{1-c} + \left(\frac{c^{j-i}}{2}\right) \frac{1-c^{n_L-j+1}}{1-c}, \quad i < j \quad (77)$$

Last we need to find the PageRank for the nodes in the complete graph not linked to by the node in the line ( $\mathbf{R}_{G,i}^{(2)}[S_L \rightarrow S_G]$ ).

We get the same probability of the node getting back to itself as for the node in the complete graph linked to by the line. As such we only need to find the sum of probability of getting to the node once after starting in all nodes once.

For the other nodes  $e_c$  in the complete graph we get:

$$P(e_c \rightarrow e_i) = \sum_{k=0}^{\infty} \frac{c}{n_G-1} \left(\frac{c(n_G-2)}{n_G-1}\right)^k \quad (78)$$

We got  $n_G - 1$  of those nodes giving:

$$\sum_{e_c \in S_G, e_c \neq e_i} P(e_c \rightarrow e_i) = \frac{c(n_G-1)}{(n_G-1) - (n_G-2)c} \quad (79)$$

For the rest of the nodes we can write the sum of probabilities of getting to  $e_g$  once from the line as:

$$P(e_{line} \rightarrow e_g) = \frac{c}{2} \frac{1-c^{n_L+1-j}}{1-c} \quad (80)$$

This gives the probability to get from the line to node  $e_i$ :

$$\begin{aligned} P(e_{line} \rightarrow e_i) &= \frac{c}{2} \frac{1-c^{n_L+1-j}}{1-c} P(e_g \rightarrow e_i) \\ &= \frac{c}{2} \frac{1-c^{n_L+1-j}}{1-c} \left(\frac{c}{(n_G-1) - (n_G-2)c}\right) = \frac{c^2(1-c^{n_L+1-j})}{2(1-c)((n_G-1) - (n_G-2)c)} \end{aligned} \quad (81)$$

Using this we get the sum of probability to get to a node in the complete graph not linked to by the line as:

$$\sum_{\substack{e_k \in S \\ e_k \neq e_i}} P(e_k \rightarrow e_i) = \frac{c^2(1-c^{n_L+1-j})}{2(1-c)((n_G-1) - (n_G-2)c)} + 1 + \frac{c(n_G-1)}{(n_G-1) - (n_G-2)c} \quad (82)$$

Multiplying with the sum of probability of going from the node in question back to itself gives the PageRank:

$$\begin{aligned} \mathbf{R}_{G,i}^{(2)}[S_L \rightarrow S_G] &= \left( \frac{((n_G - 1) - c(n_G - 2))}{((n_G - 1) - c(n_G - 2)) - c^2} \right) \\ &\left( \frac{c^2(1 - c^{n_L+1-j})}{2(1-c)((n_G - 1) - (n_G - 2)c)} + 1 + \frac{c(n_G - 1)}{(n_G - 1) - (n_G - 2)c} \right) \\ &= \left( \frac{c^2(1 - c^{n_L+1-j})}{2(1-c)} \right) \left( \frac{1}{((n_G - 1) - c(n_G - 2)) - c^2} \right) + \frac{1}{1-c} \end{aligned} \quad (83)$$

And the proof is complete  $\square$   $\square$

We include a matrix proof of Theorem 4.7 as well:

*Proof of Theorem 4.7.* We divide the matrix  $(\mathbf{I} - c\mathbf{A}^\top)$  in blocks:

$$(\mathbf{I} - c\mathbf{A}^\top) = \begin{bmatrix} \mathbf{B} & \mathbf{C} \\ \mathbf{D} & \mathbf{E} \end{bmatrix}$$

where  $\mathbf{B}$  is the part corresponding to the line,  $\mathbf{C}$  is a zero matrix (since we have no links from nodes in the complete graph to the line).  $\mathbf{D}$  is a zero matrix except for one element  $D_{g,j} = -c/2$ , where  $e_j$  is  $j$ :th the node in the line linking to the complete graph and  $e_g$  is the  $g$ :th node in the graph linked to by node  $e_j$ . We note that  $j, g$  are the internal number for the complete graph and line respectively and not their "number" in the combined graph.  $\mathbf{E}$  is the part corresponding to the complete graph.

In the same way we divide the inverse in blocks:

$$(\mathbf{I} - c\mathbf{A}^\top)^{-1} = \begin{bmatrix} \mathbf{B}^{inv} & \mathbf{C}^{inv} \\ \mathbf{D}^{inv} & \mathbf{E}^{inv} \end{bmatrix}$$

Using Lemma 4.1 for blockwise inversion we get:

$$\begin{aligned} \mathbf{B}^{inv} &= (\mathbf{B} - \mathbf{D}\mathbf{E}^{-1}\mathbf{C})^{-1} = \mathbf{B}^{-1} \\ \mathbf{C}^{inv} &= -(\mathbf{B} - \mathbf{D}\mathbf{E}^{-1}\mathbf{C})^{-1}\mathbf{C}\mathbf{E}^{-1} = 0 \\ \mathbf{D}^{inv} &= -\mathbf{E}^{-1}\mathbf{D}(\mathbf{B} - \mathbf{D}\mathbf{E}^{-1}\mathbf{C})^{-1} = \mathbf{E}^{-1}\mathbf{D}\mathbf{B}^{-1} \\ \mathbf{E}^{inv} &= \mathbf{E}^{-1} + \mathbf{E}^{-1}\mathbf{D}(\mathbf{B} - \mathbf{D}\mathbf{E}^{-1}\mathbf{C})^{-1}\mathbf{C}\mathbf{E}^{-1} = \mathbf{E}^{-1} \end{aligned} \quad (84)$$

Since one of the nodes in the line links out we get  $\mathbf{B}$  divided in blocks:

$$\mathbf{B} = \begin{bmatrix} \mathbf{B}_B & \mathbf{B}_C \\ \mathbf{B}_D & \mathbf{B}_E \end{bmatrix}$$

$$\mathbf{B}_B = \begin{bmatrix} 1 & -c & 0 & \dots & 0 \\ 0 & 1 & -c & \ddots & \vdots \\ 0 & 0 & 1 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -c \\ 0 & \dots & 0 & 0 & 1 \end{bmatrix} \quad \mathbf{B}_C = \begin{bmatrix} 0 & \dots & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \vdots \\ -c/2 & 0 & \dots & 0 \end{bmatrix}$$



where  $\mathbf{B}_D$  is a zero matrix and  $\mathbf{B}_E$  looks the same as  $\mathbf{B}_B$  although possibly with a different size. The size of the blocks are:  $\mathbf{B}_B : (j-1) \times (j-1)$ ,  $\mathbf{B}_C : (j-i) \times (n_L - j + 1)$ ,  $\mathbf{B}_D : (n_L - j + 1) \times (n_L - j + 1)$  and  $\mathbf{B}_E : (n_L - j + 1) \times (n_L - j + 1)$ , where  $n_L$  is the total number of nodes in the line.

For the blocks of the inverse we get:

$$\begin{aligned}\mathbf{B}_B^{inv} &= \mathbf{B}_B^{-1} \\ \mathbf{B}_C^{inv} &= -\mathbf{B}_B^{-1} \mathbf{B}_C \mathbf{B}_E^{-1} \\ \mathbf{B}_D^{inv} &= 0 \\ \mathbf{B}_E^{inv} &= \mathbf{E}^{-1}\end{aligned}\tag{85}$$

$\mathbf{B}_B^{inv}$  and  $\mathbf{B}_E^{inv}$  are found as the inverse for the simple line, leaving  $\mathbf{B}_C^{inv}$  to be computed. The only difference compared to a simple line is that the only non-zero element in  $\mathbf{B}_C$  is  $-c/2$  rather than  $-c$ . In other words  $\mathbf{B}^{-1}$  is exactly as it would have been for a simple line, except block corresponding to  $\mathbf{B}_C^{inv}$  which is multiplied with 0.5.

We can now find the PageRank of the nodes in the line:

$$\begin{aligned}\mathbf{R}_{L,i}^{(2)}[S_L \rightarrow S_G] &= \sum_{k=1}^{n_L} B_{i,k}^{-1} = \sum_{k=1}^{j-1} B_{i,k}^{-1} + \sum_{k=j}^{n_L} B_{i,k}^{-1} \\ &= \sum_{k=0}^{j-i-1} c^k + \sum_{k=j-i}^{n_L-i} \frac{c^k}{2} = \frac{1 - c^{j-i}}{1 - c} + \left( \frac{c^{j-i}}{2} \right) \frac{1 - c^{n_L-j+1}}{1 - c}\end{aligned}\tag{86}$$

For the nodes in the complete graph we first need to find  $\mathbf{D}^{inv}$ , to do so we start by calculating  $\mathbf{D}\mathbf{B}^{-1}$ . Since only one element  $D_{gj}$  of  $\mathbf{D}$  is non-zero, only row  $g$  of  $\mathbf{D}\mathbf{B}^{-1}$  can be non-zero. We get row  $g$  as:

$$(\mathbf{D}\mathbf{B}^{-1})_{\text{row}_g} = \frac{-c}{2} [0 \ \dots \ 1 \ c \ c^2 \ \dots \ c^L - j]$$

where there are  $j-1$  zeros before the 1, ( $\mathbf{B}^{-1}$  upper triangular). Multiplying this with  $\mathbf{E}^{-1}$  found in the matrix proof of Lemma 4.2 gives:

$$\begin{aligned}-\mathbf{E}^{-1}\mathbf{D}\mathbf{B}^{-1} &= \frac{c}{2} \begin{bmatrix} 0 & \dots & 0 & s & cs & \dots & c^{n_L-j}s \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & s & cs & \dots & c^{n_L-j}s \\ 0 & \dots & 0 & d & cd & \dots & c^{n_L-j}d \\ 0 & \dots & 0 & s & cs & \dots & c^{n_L-j}s \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & s & cs & \dots & c^{n_L-j}s \end{bmatrix} \\ s &= \frac{c}{(n_G - 1) - c(n_G - 2) - c^2}, \quad d = \frac{(n_G - 1) - c(n_G - 2)}{(n_G - 1) - c(n_G - 2) - c^2}\end{aligned}$$

We can now find the PageRank of the nodes in the complete graph by summation of corresponding row:

$$\mathbf{R}_{G,i}^{(2)}[S_L \rightarrow S_G] = \sum_{k=1}^{n_L} D_{ik}^{inv} + \sum_{k=1}^{n_L} E_{ik}^{inv} = \sum_{k=1}^{n_L} D_{ik}^{inv} + \frac{1}{1-c} \quad (87)$$

We separate between the node in the complete graph linked to from the line and the other nodes in the complete graph.

$$\begin{aligned} \mathbf{R}_{G,i}^{(2)}[S_L \rightarrow S_G] &= \frac{c}{2} \sum_{k=0}^{n_L-j} c^k s + \frac{1}{1-c} \\ &= \left( \frac{c(1-c^{n_L-j+1})}{2(1-c)} \right) \left( \frac{c}{(n_G-1)-c(n_G-2)-c^2} \right) + \frac{1}{1-c}, \quad i \neq g \end{aligned} \quad (88)$$

$$\begin{aligned} \mathbf{R}_{G,g}^{(2)}[S_L \rightarrow S_G] &= \frac{c}{2} \sum_{k=0}^{n_L-j} c^k d + \frac{1}{1-c} \\ &= \left( \frac{c(1-c^{n_L+1-j})}{2(1-c)} \right) \left( \frac{((n_G-1)-c(n_G-2))}{((n_G-1)-c(n_G-2))-c^2} \right) + \frac{1}{1-c} \end{aligned} \quad (89)$$

And the proof is complete. For completeness we include the complete inverse as well:

$$\begin{aligned} (\mathbf{I} - c\mathbf{A}^\top)^{-1} &= \begin{bmatrix} \mathbf{B}^{inv} & \mathbf{C}^{inv} \\ \mathbf{D}^{inv} & \mathbf{E}^{inv} \end{bmatrix}, \quad \mathbf{B}^{-1} = \begin{bmatrix} 1 & c & c^2 & \dots & c^{n_L-1}/2 \\ 0 & 1 & c & \dots & c^{n_L-2}/2 \\ 0 & 0 & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & c/2 \\ 0 & 0 & \dots & 0 & 1 \end{bmatrix} \\ \mathbf{C}^{-1} &= \begin{bmatrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{bmatrix}, \quad \mathbf{D}^{-1} = -\mathbf{E}^{-1}\mathbf{D}\mathbf{B}^{-1} \text{ (seen above)} \\ \mathbf{E}^{-1} &= \begin{bmatrix} 1 - (n_G-1)ak & k & k & \dots & k \\ k & 1 - (n_G-1)ak & k & \dots & k \\ k & k & 1 - (n_G-1)ak & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & k \\ k & k & \dots & k & 1 - (n_G-1)ak \end{bmatrix} \\ a &= -c/(n-1), \quad k = \frac{-a}{(1-a^2 + (n-2)(a-a^2))} \end{aligned}$$

□

□

The normalizing constant can then be found by summation of the individual PageRank of all the nodes in order to get the normalized PageRank  $\mathbf{R}^{(1)}$ .

We note that while the node in the line that links to the complete graph does not lose anything from the new link, the nodes below it in the line do lose quite a lot because of it. Likewise the PageRank of the node that's linked to gains more from the new link than the others in the complete graph.

#### 4.3.3 Connecting the simple line with a complete graph by letting one node in the line be part of the complete graph

If we instead let one node in the line be part of the complete graph we get another interesting example to look at. An example of what the system could look like can be seen in Fig. 7.

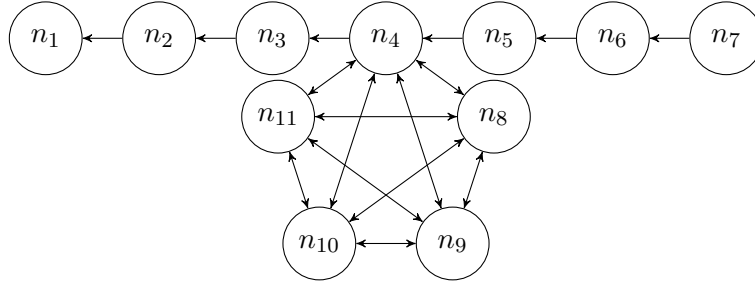


Figure 7: Simple line with one node in the line being a part of a complete graph

We formulate the following theorem for the PageRank of the given example:

**Theorem 4.8.** *The PageRank of the nodes in system  $S_L$  made up of a simple line and system  $S_G$  made up of a complete graph after one of the nodes  $e_j \in S_L$  becomes part of the complete graph assuming uniform  $\mathbf{u}$  can be written:*

$$\mathbf{R}_{L,i}^{(2)}[S_L \leftrightarrow S_G] = \frac{1 - c^{n_L+1-i}}{1 - c}, \quad i > j \quad (90)$$

$$\mathbf{R}_{L,j}^{(2)}[S_L \leftrightarrow S_G] = \left( \frac{1 - c^{n_L+1-j}}{1 - c} + \frac{c(n_G - 1)}{(n_G - 1) - c(n_G - 2)} \right) \left( \frac{n_G((n_G - 1) - c(n_G - 2))}{n_G((n_G - 1) - c(n_G - 2)) - c^2(n_G - 1)} \right) \quad (91)$$

$$\mathbf{R}_{L,i}^{(2)}[S_L \leftrightarrow S_G] = \frac{c^{j-i} \mathbf{R}_{L,j}^{(2)}}{n_G} + \frac{1 - c^{j-i}}{1 - c}, \quad i < j \quad (92)$$

$$\mathbf{R}_{G,i}^{(2)}[S_L \leftrightarrow S_G] = \frac{(c + n_G)(n_G - 1)(1 - c) + (n_G - 1)c^2(1 - c^{n_L-j})}{(1 - c)(n_G(n_G - 1) - (n_G - 1)c^2 - n_G(n_G - 2)c)} \quad (93)$$

where  $\mathbf{R}_{G,i}^{(2)}[S_L \leftrightarrow S_G]$  is the PageRank for the nodes in the complete graph (except the node also being a part of the line) and  $\mathbf{R}_{L,i}^{(2)}[S_L \leftrightarrow S_G]$  is the PageRank of nodes in the line.  $n_G, n_L$  is the number of nodes in the complete graph and simple line respectively after making one node in the line part of the complete graph.

*Proof.* For the proof of the nodes  $e_i \in S_L$ ,  $i > j$  we get the PageRank for a simple line. In order to find  $\mathbf{R}_{L,j}^{(2)}[S_L \leftrightarrow S_G]$  we first use Theorem 4.1 to write it as:

$$\mathbf{R}_{L,j}^{(2)}[S_L \leftrightarrow S_G] = \left( \sum_{e_i \in S, e_i \neq e_j} P(e_i \rightarrow e_j) + 1 \right) \left( \sum_{k=0}^{\infty} (P(e_j \rightarrow e_j))^k \right) \quad (94)$$

where  $P(e_i \rightarrow e_j)$  is the probability of getting from node  $e_i$  to node  $e_j$ .

$$\begin{aligned} \sum_{e_i \in S, e_i \neq e_j} P(e_i \rightarrow e_j) + 1 &= \sum_{k=0}^{n_L+1-j} c^k + (n-1) \frac{c}{n_G-1} \sum_{k=0}^{\infty} \left( \frac{c(n_G-2)}{(n_G-1)} \right)^k \\ &= \frac{1 - c^{n_L+1-j}}{1-c} + \frac{c(n_G-1)}{(n_G-1) - c(n_G-2)} \end{aligned} \quad (95)$$

$$\begin{aligned} P(e_j \rightarrow e_j) &= \frac{c(n_G-1)}{n_G} \frac{c}{n_G-1} + \frac{c(n_G-1)}{n_G} \frac{c(n_G-2)}{n_G-1} \frac{c}{n_G-1} \\ &\quad + \frac{c(n_G-1)}{n_G} \frac{c^2(n_G-2)^2}{(n_G-1)^2} \frac{c}{n_G-1} + \dots \\ &= \frac{c^2}{n_G} \sum_{k=0}^{\infty} \left( \frac{c(n_G-2)}{n-1} \right)^k = \frac{c^2(n_G-1)}{n_G((n_G-1) - c(n_G-2))} \end{aligned} \quad (96)$$

$$\sum_{k=0}^{\infty} (P(e_j \rightarrow e_j))^k = \frac{n_G((n_G-1) - c(n_G-2))}{n_G((n_G-1) - c(n_G-2)) - c^2(n_G-1)} \quad (97)$$

Multiplication of the results from equation (95) and (97) gives

$$\begin{aligned} \mathbf{R}_{L,j}^{(2)}[S_L \leftrightarrow S_G] &= \left( \frac{1 - c^{n_L+1-j}}{1-c} + \frac{c(n_G-1)}{(n_G-1) - c(n_G-2)} \right) \\ &\quad \left( \frac{n_G((n_G-1) - c(n_G-2))}{n_G((n_G-1) - c(n_G-2)) - c^2(n_G-1)} \right) \end{aligned} \quad (98)$$

For the nodes below the one in the complete graph we can write the PageRank as:

$$\mathbf{R}_{L,i}^{(2)}[S_L \leftrightarrow S_G] = \mathbf{R}_{L,j}^{(2)}[S_L \leftrightarrow S_G] P(e_j \rightarrow e_i) + \sum_{k=0}^{j-i-1} c^k, \quad i < j \quad (99)$$

$$P(e_j \rightarrow e_i) = \frac{c^{j-i}}{n_G} \quad (100)$$

This gives:

$$\mathbf{R}_{L,i}^{(2)}[S_L \leftrightarrow S_G] = \frac{c^{j-i} \mathbf{R}_{L,j}^{(2)}[S_L \leftrightarrow S_G]}{n_G} + \frac{1 - c^{j-i}}{1 - c}, i < j \quad (101)$$

Left to prove we have the formula for the nodes in the complete graph not directly connected to the line  $\mathbf{R}_{G,i}^{(2)}[S_L \leftrightarrow S_G]$ . We do not need to consider the part of the line following the complete graph, since we can not get from this part of the graph back to the complete graph. We already have the PageRank for the nodes not linking out in the complete graph in the case where we have no line of nodes linking to the complete graph from Theorem 4.4. In the Matrix proof there we found the PageRank of the nodes in the complete graph not linking out to be:

$$\begin{aligned} \mathbf{R}_{G,i}^{(2)}[S_G] &= D_i^{inv} + E_d^{inv} + (n_G - 2)E_e^{inv} \\ &= \left( P(e_j \rightarrow e_i) + 1 + \sum_{\substack{e_k \in S_G \\ e_k \neq e_j, e_i}} P(e_k \rightarrow e_i) \right) \left( \sum_{k=0}^{\infty} (P(e_i \rightarrow e_i))^k \right) \end{aligned} \quad (102)$$

We identify  $D_i^{inv}$  in the expression:

$$D_i^{inv} = P(e_j \rightarrow e_i) \left( \sum_{k=0}^{\infty} (P(e_i \rightarrow e_i))^k \right) \quad (103)$$

Since all paths  $P(e_k \rightarrow e_i)$  where  $e_k \in S_L$  need to go through node  $e_j$  we can write these as a product of the probability to get to node  $e_j$  times the probability to from there get to node  $e_i$  for which we want to calculate PageRank.

$$\begin{aligned} \mathbf{R}_{G,i}^{(2)}[S_L \leftrightarrow S_G] &= \left( \sum_{k=0}^{\infty} (P(e_i \rightarrow e_i))^k \right) \\ &\left( \left( \sum_{\substack{e_k \in S_L \\ e_k \neq e_j}} P(e_k \rightarrow e_j) + 1 \right) P(e_j \rightarrow e_i) + 1 + \sum_{\substack{e_k \in S_G \\ e_k \neq e_j, e_i}} P(e_k \rightarrow e_i) \right) \end{aligned} \quad (104)$$

Using the expressions found in the matrix proof of Theorem 4.4 we get PageRank.

$$\begin{aligned} \mathbf{R}_{G,i}^{(2)}[S_L \leftrightarrow S_G] &= \left( \sum_{k=0}^{n_L-j} c^k \right) D_i^{inv} + E_d^{inv} + (n_G - 2)E_e^{inv} \\ &= \frac{(n_G - 1)(n_G + c)}{n_G(n_G - 1) - n_G(n_G - 2)c - (n_G - 1)c^2} \\ &+ \left( \frac{c(1 - c^{n_L-j})}{1 - c} \right) \frac{c(n_G - 1)}{n_G(n_G - 1) - n_G(n_G - 2)c - (n_G - 1)c^2} \\ &= \frac{(c + n_G)(n_G - 1)(1 - c) + (n_G - 1)c^2(1 - c^{n_L-j})}{(1 - c)(n_G(n_G - 1) - (n_G - 1)c^2 - n_G(n_G - 2)c)} \end{aligned} \quad (105)$$

□

□

We give a matrix proof of Theorem 4.8 as well.

*Proof of Theorem 4.8.* For the proof we consider the structure in Fig. 7 when we talk about nodes below/above other nodes.

For the proof we once again start by dividing the matrix  $(I - cA^\top)$  in blocks:

$$(I - cA^\top) = \begin{bmatrix} B & C \\ D & E \end{bmatrix}$$

And the same for the inverse:

$$(I - cA^\top)^{-1} = \begin{bmatrix} B^{inv} & C^{inv} \\ D^{inv} & E^{inv} \end{bmatrix}$$

We let  $B$  be the matrix for the nodes "below" the complete graph as well as the complete graph. In other words we can divide  $B$  itself in blocks:

$$B = \begin{bmatrix} B_B & B_C \\ B_D & B_E \end{bmatrix}$$

where  $B_B$  corresponds to the part of the line we can reach from the complete graph (nodes below the complete graph),  $B_C$  is a zero matrix except for the bottom left element which is equal to  $-c/n_G$  corresponding to the link from the complete graph to the linked to node in the line.  $B_D$  is a zero matrix and  $B_E$  corresponds to the matrix of a complete graph with the first node in the complete graph linking out (any other could of course be used and would result in the same calculations after multiplication with a suitable permutation matrix.).

Then  $B$  have the same structure as the matrix for our example system with one node in a complete graph linking to a node in the line in Fig. 5 (except we have no nodes "above" the complete graph). We note that we have already looked at this system in the matrix proof of Theorem 4.6 and will leave it until we will need it later.

We let  $E$  correspond to the part of the line "above" the complete graph, which is also something we have looked at multiple times already.  $C$  is a zero matrix except for element  $C_{j1}$  corresponding to the link from the line to the complete graph.  $D$  is a zero matrix.

For completeness we include the sizes of the different blocks as well,  $B : (j-1+n_G) \times (j-1+n_G)$ ,  $C : (j-1+n_G) \times (n_L-j)$ ,  $D : (n_L-j) \times (j-1+n_G)$ ,  $E : (n_L-j) \times (n_L-j)$ .

From Lemma 4.1 we get:

$$\begin{aligned} B^{inv} &= (B - DE^{-1}C)^{-1} = B^{-1} \\ C^{inv} &= -(B - DE^{-1}C)^{-1}CE^{-1} = B^{-1}CE^{-1} \\ D^{inv} &= -E^{-1}D(B - DE^{-1}C)^{-1} = 0 \\ E^{inv} &= E^{-1} + E^{-1}D(B - DE^{-1}C)^{-1}CE^{-1} = E^{-1} \end{aligned} \tag{106}$$

We already know most of  $B^{-1}$  from the matrix proof of Theorem 5 and  $E^{-1}$  from the matrix proof of Theorem 4.2. Left to find is  $C^{inv}$  before we can find the PageRank.

Calculating  $\mathbf{C}\mathbf{E}^{-1}$  we get a  $(j - 1 + n_G) \times (n_L - j)$  zero matrix except for row  $j$  corresponding to the non-zero element  $C_{j1} = -c$  in  $\mathbf{C}$ . This gives  $\mathbf{C}\mathbf{E}^{-1}$ :

$$\mathbf{C}\mathbf{E}^{-1} = -c \begin{bmatrix} 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & 0 & \dots & 0 \\ 1 & c & c^2 & \dots & c^{n_L-j-1} \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}$$

where row  $j$  is the non-zero row. When multiplying this with  $\mathbf{B}^{-1}$  we are only interested in column  $j$  of  $\mathbf{B}^{-1}$  since only row  $j$  of  $\mathbf{C}\mathbf{E}^{-1}$  is non-zero. From the matrix proof of Theorem 5 we get column  $j$  as:

$$B_{kj}^{-1} = \begin{cases} \frac{c^{j-k}}{n_G} (1 - (n_G - 1)bk_A), & 1 \leq k < j \\ (1 - (n_G - 1)bk_A), & k = j \\ \frac{bk_A}{a}, & j < k \leq j - 1 + n_G \end{cases} \quad (107)$$

$$k_A = \frac{-a}{(1 - ab) + (n_G - 2)(a - ab)}$$

$$a = \frac{-c}{n_G - 1}, \quad b = \frac{-c}{n_G}$$

We note that we get  $c^{j-k}$  rather than  $c^{j-k+1}$  since node  $e_j$  in this case is the node linking to node  $e_{j-1}$  rather than node  $e_j$  being the one linked to. We are now ready to calculate  $\mathbf{C}^{inv}$ :

$$\mathbf{C}_{\text{row}_k}^{inv} = B_{kj}^{-1} [c \ c^2 \ \dots \ c^{n_L-j}], \quad (1 \leq k \leq j - 1 + n_G) \quad (108)$$

To get the PageRank of a node we need to sum all the elements of corresponding row of  $(\mathbf{I} - c\mathbf{A}^\top)^{-1}$ , for the nodes "above" the complete graph we get the simple line:

$$\mathbf{R}_{L,i}^{(2)}[S_L \leftrightarrow S_G] = \sum_{k=0}^{n_L-i} c^k = \frac{1 - c^{n_L-i+1}}{1 - c}, \quad i > j \quad (109)$$

For the node part of both the complete graph and the line we get:

$$\begin{aligned}
\mathbf{R}_{L,j}^{(2)}[S_L \leftrightarrow S_G] &= \sum_{k=1}^{n_G} B_{jk}^{-1} + \sum_{k=1}^{n_L-j} C_{jk}^{inv} \\
&= \frac{n_G(n_G-1) + n_G c}{n_G(n_G-1) - (n_G-1)c^2 - n_G(n_G-2)c} + B_{jj}^{-1} \sum_{k=1}^{n_L-j} c^k \\
&= \frac{n_G(n_G-1) + n_G c}{n_G(n_G-1) - (n_G-1)c^2 - n_G(n_G-2)c} \\
&+ \left( \frac{n_G(n_G-1) - n_G(n_G-2)c}{n_G(n_G-1) - n_G(n_G-2)c - 2(n_G-2)c^2} \right) \left( \sum_{k=0}^{n_L-j} c^k - 1 \right) \\
&= \left( \frac{1 - c^{n_L+1-j}}{1-c} + \frac{c(n_G-1)}{(n_G-1) - c(n_G-2)} \right) \\
&\left( \frac{n_G((n_G-1) - c(n_G-2))}{n_G((n_G-1) - c(n_G-2)) - c^2(n_G-1)} \right)
\end{aligned} \tag{110}$$

For the nodes "below" the complete graph we get:

$$\begin{aligned}
\mathbf{R}_{L,i}^{(2)}[S_L \leftrightarrow S_G] &= \sum_{k=1}^{n_G} B_{ik}^{-1} + \sum_{k=1}^{n_L-j} C_{ik}^{inv}, \quad i < j \\
&= \sum_{k=0}^{j-i-1} c^k + \frac{c^{j-i}(c + (n_G-1))}{n_G(n_G-1) - n_G(n_G-2)c - (n_G-1)c^2} + B_{ij}^{-1} \sum_{k=1}^{n_L-j} c^k
\end{aligned} \tag{111}$$

where we once again note that we get  $c^{j-i}$  rather than  $c^{j-i+1}$  since we consider node  $j-1$  the node linked to by the graph rather than node  $j$ .

$$\begin{aligned}
\mathbf{R}_{L,i}^{(2)}[S_L \leftrightarrow S_G] &= \frac{1 - c^{j-i}}{1-c} \\
&+ \frac{c^{j-i}(c + (n_G-1))}{n_G(n_G-1) - n_G(n_G-2)c - (n_G-1)c^2} + \frac{c^{j-i}}{n_G} B_{jj}^{-1} \sum_{k=1}^{n_L-j} c^k \\
&= \frac{1 - c^{j-i}}{1-c} + \frac{c^{j-i}}{n_G} \mathbf{R}_{L,j}^{(2)}[S_L \leftrightarrow S_G]
\end{aligned} \tag{112}$$

For the nodes in the complete graph not part of the line we get:

$$\mathbf{R}_{G,i}^{(2)}[S_L \leftrightarrow S_G] = \sum_{k=1}^{n_G} B_{ik}^{-1} + \sum_{k=1}^{n_L-j} C_{jk}^{inv}, \quad i \neq j \tag{113}$$



where  $e_j \in S_G$  is the node in the complete graph also part of the line.

$$\begin{aligned}
\mathbf{R}_{G,i}^{(2)} &= \frac{(c + n_G)(n_G - 1)}{n_G(n_G - 1) - (n_G - 1)c^2 - n_G(n_G - 2)c} + \frac{bk_A}{a} \sum_{k=1}^{n_L-j} c^k \\
&= \frac{(c + n_G)(n_G - 1)}{n_G(n_G - 1) - (n_G - 1)c^2 - n_G(n_G - 2)c} \\
&\quad + \frac{(n_G - 1)c^2(1 - c^{n_L-j})}{(1 - c)(n_G(n_G - 1) - (n_G - 1)c^2 - n_G(n_G - 2)c)} \\
&= \frac{(c + n_G)(n_G - 1)(1 - c) + (n_G - 1)c^2(1 - c^{n_L-j})}{(1 - c)(n_G(n_G - 1) - (n_G - 1)c^2 - n_G(n_G - 2)c)}
\end{aligned} \tag{114}$$

And the proof is complete.  $\square$   $\square$

For reference we include the complete inverse matrix once again.

$$\begin{aligned}
(\mathbf{I} - c\mathbf{A}^\top)^{-1} &= \begin{bmatrix} \mathbf{B}_B^{inv} & \mathbf{B}_C^{inv} & \mathbf{C}_1^{inv} \\ \mathbf{B}_D^{inv} & \mathbf{B}_E^{inv} & \mathbf{C}_2^{inv} \\ \mathbf{D}_1^{inv} & \mathbf{D}_2^{inv} & \mathbf{E}^{inv} \end{bmatrix} \\
\mathbf{B}_B^{inv} &= \begin{bmatrix} 1 & c & c^2 & \dots & c^{j-2} \\ 0 & 1 & c & \dots & c^{j-3} \\ 0 & 0 & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & c \\ 0 & 0 & \dots & 0 & 1 \end{bmatrix}, \quad \mathbf{B}_C^{inv} = \begin{bmatrix} \frac{c^{j-1}s_A}{n_G} & \frac{c^{j-1}k_A}{c^{j-2}s_A} & \frac{c^{j-1}k_A}{c^{j-2}k_A} & \dots & \frac{c^{j-1}k_A}{c^{j-2}k_A} \\ \frac{n_G}{c^{j-2}s_A} & \frac{n_G}{c^{j-2}k_A} & \frac{n_G}{c^{j-2}k_A} & \dots & \frac{n_G}{c^{j-2}k_A} \\ \frac{\vdots}{n_G} & \frac{\vdots}{n_G} & \frac{\vdots}{n_G} & \dots & \frac{\vdots}{n_G} \\ \frac{cs_A}{n_G} & \frac{ck_A}{n_G} & \frac{ck_A}{n_G} & \dots & \frac{ck_A}{n_G} \end{bmatrix} \\
\mathbf{B}_D^{inv} &= \begin{bmatrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{bmatrix}, \quad \mathbf{B}_E^{inv} = \begin{bmatrix} \frac{s_A}{bk_A} & k_A & k_A & \dots & k_A \\ \frac{a}{bk_A} & s_D & k_D & \dots & k_D \\ \frac{bk_A}{a} & k_D & s_D & \ddots & k_D \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \frac{bk_A}{a} & k_D & k_D & \dots & s_D \end{bmatrix} \\
s_A &= 1 - (n_G - 1)bk_A, \quad s_D = 1 - (n_G - 2)ak_D
\end{aligned}$$

Where  $\mathbf{B}$  is taken from the matrix proof of Theorem 4.6 where  $\mathbf{B}$  is the part of the graph consisting of a line with  $j - 1$  nodes and a complete graph with  $n_G$  nodes with the first

node in the complete graph linking to node  $j - 1$  in the line.

$$\mathbf{C}^{inv} = \begin{bmatrix} \mathbf{C}_1^{inv} \\ \mathbf{C}_2^{inv} \end{bmatrix} = \begin{bmatrix} \frac{-c^{j-1}cs_A}{n_G} & \frac{-c^{j-1}c^2s_A}{n_G} & \dots & \frac{-c^{j-1}c^{n_L-j}s_A}{n_G} \\ \frac{-c^{j-2}cs_A}{n_G} & \frac{-c^{j-2}c^2s_A}{n_G} & \dots & \frac{-c^{j-2}c^{n_L-j}s_A}{n_G} \\ \vdots & \vdots & \dots & \vdots \\ \frac{-c^1cs_A}{n_G} & \frac{-c^1c^2s_A}{n_G} & \dots & \frac{-c^1c^{n_L-j}s_A}{n_G} \\ \frac{-cs_A}{n_G} & \frac{-c^2s_A}{n_G} & \dots & \frac{-c^{n_L-j}s_A}{n_G} \\ \frac{-cbk_A}{n_G} & \frac{-c^2bk_A}{n_G} & \dots & \frac{-c^{n_L-j}bk_A}{n_G} \\ \frac{a}{-cbk_A} & \frac{a}{-c^2bk_A} & \dots & \frac{a}{-c^{n_L-j}bk_A} \\ \vdots & \vdots & \dots & \vdots \\ \frac{-cbk_A}{a} & \frac{-c^2bk_A}{a} & \dots & \frac{-c^{n_L-j}bk_A}{a} \end{bmatrix}$$

$$\mathbf{D}^{inv} = [\mathbf{D}_1^{inv} \ \mathbf{D}_2^{inv}] = \begin{bmatrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{bmatrix}, \quad \mathbf{E}^{inv} = \begin{bmatrix} 1 & c & c^2 & \dots & c^{n_L-j-1} \\ 0 & 1 & c & \dots & c^{n_L-j-2} \\ 0 & 0 & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & c \\ 0 & 0 & \dots & 0 & 1 \end{bmatrix}$$

**Theorem 4.9.** *The normalizing constant  $N$  for the simple line with one node being part of a complete graph using uniform  $\mathbf{u}$  can be written as:*

$$N = (n_G - 1) \mathbf{R}_{G,i \notin L}^{(2)}[S_L \leftrightarrow S_G] + \mathbf{R}_{L,j}^{(2)}[S_L \leftrightarrow S_G] + \frac{n_L - 1}{1 - c} - \frac{c(1 - c^{n_L-j})}{(1 - c)^2} - \frac{c(1 - c^{j-1})}{(1 - c)^2} + \frac{c(1 - c^{j-1}) \mathbf{R}_{L,j}^{(2)}[S_L \leftrightarrow S_G]}{n_G(1 - c)} \quad (115)$$

where  $\mathbf{R}_{G,i}^{(2)}[S_L \leftrightarrow S_G]$  is the PageRank of nodes in the complete graph,

$\mathbf{R}_{L,j}^{(2)}[S_L \leftrightarrow S_G]$  is for the node in both the line and complete graph and

$\mathbf{R}_{L,i}^{(2)}[S_L \leftrightarrow S_G]$  is for the nodes in the line.

*Proof.* The normalizing constant is equal to the sum of the non-normalized PageRank of all nodes.

We got  $n_G$  nodes in the complete graph,  $(n - 1)$  not directly connected to the line and one connected to the line. This gives:

$$N = (n - 1) \mathbf{R}_{G,i}^{(2)}[S_L \leftrightarrow S_G] + \mathbf{R}_{L,j}^{(2)}[S_L \leftrightarrow S_G] + \sum_{i \neq j} \mathbf{R}_{L,j}^{(2)}[S_L \leftrightarrow S_G] \quad (116)$$

where  $\mathbf{R}_{L,j}^{(2)}[S_L \leftrightarrow S_G]$  is the PageRank of individual nodes in the line except for the node node  $j$  in the line for which we have  $\mathbf{R}_{L,j}^{(2)}[S_L \leftrightarrow S_G]$ . For those nodes we got PageRank:

$$\mathbf{R}_{L,j}^{(2)}[S_L \leftrightarrow S_G] = \begin{cases} \frac{1 - c^{n_L+1-i}}{1 - c}, & i > j \\ \frac{c^{j-i} \mathbf{R}_{L,j}^{(2)} + \frac{1 - c^{j-i}}{1 - c}}{n_G}, & i < j \end{cases} \quad (117)$$

The sum of all nodes for which  $i > j$  can be written:

$$\sum_{i=j+1}^{n_L} \mathbf{R}_{L,j}^{(2)}[S_L \leftrightarrow S_G] = \frac{n_L - j}{1 - c} - \frac{c(1 - c^{n_L-j})}{(1 - c)^2} \quad (118)$$

where we use that the second part  $\sum_{i=j}^{n_L} \frac{-c^{n_L+1-i}}{1 - c}$  is a geometric sum. Calculating the sum for  $i < j$  in the same way we get:

$$\sum_{i=1}^{j-1} \mathbf{R}_{L,j}^{(2)}[S_L \leftrightarrow S_G] = \frac{j-1}{1 - c} - \frac{c(1 - c^{j-1})}{(1 - c)^2} + \frac{c(1 - c^{j-1}) \mathbf{R}_{L,j}^{(2)}}{n_G(1 - c)} \quad (119)$$

Summation of all individual parts completes the proof.  $\square$

Now that we have an explicit formula for this example we can look at what happens when we change various parameters like  $c$  or the size of either the line or complete graph.

#### 4.4 A closer look at the formulas for PageRank in our examples

Now that we have formulas for the PageRank of a couple different graph structures we are going to take a short look at what happens when we change some parameters. We will also take a look at the partial derivative with respect to  $c$ .

##### 4.4.1 Partial derivatives with respect to $c$

In the case of the simple line with formula as seen earlier we get the derivative with respect to  $c$  as:

$$\frac{\partial}{\partial c} \mathbf{R}_i^{(2)}[S_L] = (1 - c)^{-2} - \frac{c^{n_L-i+1} (n_L - i + 1)}{c(1 - c)} - \frac{c^{n_L-i+1}}{(1 - c)^2} \quad (120)$$

Rewriting it and looking to see if it is positive we get:

$$\frac{1 + c^{n_L-i} (i - n_L)(1 - c)}{(-1 + c)^2} \geq 0 \Leftrightarrow c^{n_L-i} ((i - n_L)(1 - c) + \frac{1}{c^{n_L-i}}) \geq 0 \quad (121)$$

$$\Leftrightarrow \frac{1}{c^{n_L-i}} \geq (n_L-i)(1-c) \Leftrightarrow \frac{1}{1-c} \geq (n_L-i)c^{n_L-i} \Leftrightarrow \sum_{k=0}^{\infty} c^k \geq (n_L-i)c^{n_L-i} \quad (122)$$

Since we have  $0 < c < 1$ ,  $n_L \geq i$  we have that  $c^k > c^{k+1}$  the first  $n_L - i$  elements of the left sum is at least as large as  $c^{n_L-i}$ , this gives:

$$\sum_{k=0}^{\infty} c^k \geq \sum_{k=1}^{n_L-i} c^k \geq (n_L-i)c^{n_L-i} \quad (123)$$

For our case with a line connected to a complete graph by letting one node in the complete graph be part of the line we get the following derivative with respect to  $c$ :

$$\begin{aligned} & \frac{\partial}{\partial c} \mathbf{R}_{L,j}^{(2)}[S_L \leftrightarrow S_G] \\ &= \frac{\left( \left( (-1+c)n_G^2 + (-1+c)^2 n_G - c^2 \right) \left( (-1+c)n_G - 2c + 1 \right) \frac{\partial}{\partial c} G(c) \right) n_G}{\left( (-1+c)n_G^2 + (-1+c)^2 n_G - c^2 \right)^2} \\ & \quad - \frac{\left( (n_G-1) \left( c \left( (-2+c)n_G + 2 - 2c \right) G(c) - (n_G-1)(n_G+c^2) \right) \right) n_G}{\left( (-1+c)n_G^2 + (-1+c)^2 n_G - c^2 \right)^2} \\ & \quad G(c) = \frac{1 - c^{n_L-j+1}}{1-c} \\ & \quad \frac{\partial}{\partial c} G(c) = (1-c)^{-2} - \frac{c^{n_L-j+1} (n_L-j+1)}{c(1-c)} - \frac{c^{L-j+1}}{(1-c)^2} \end{aligned} \quad (124)$$

The derivative have about the same shape as the original function. As  $c$  gets large so does the derivative and as  $n_G$  increases the slope get steeper for large  $c$ .

Looking at the other nodes in the complete graph we get the derivative with respect to  $c$  as:

$$\begin{aligned} & \frac{\partial}{\partial c} \mathbf{R}_{G,i}^{(2)}[S_L \leftrightarrow S_G] \\ &= \frac{(n_G-1)(1-c) - (c-n_G)(n_G-1) + 2(n_G-1)c(1-c^{n_L-j})}{(1-c)(n_G(n_G-1) - (n_G-1)c^2 - n_G(n_G-2)c)} \\ & \quad - \frac{(n_G-1)c^{1+n_L-j}(n_L-j)}{(1-c)(n_G(n_G-1) - (n_G-1)c^2 - n_G(n_G-2)c)} \\ & \quad + \frac{(c+n_G)(n_G-1)(1-c) + (n_G-1)c^2(1-c^{n_L-j})}{(1-c)^2(n_G(n_G-1) - (n_G-1)c^2 - n_G(n_G-2)c)} \\ & \quad - \frac{((c+n_G)(n_G-1)(1-c) + (n_G-1)c^2(1-c^{n_L-j}))(2c + (2-2c-n_G)n_G)}{(1-c)(n_G(n_G-1) - (n_G-1)c^2 - n_G(n_G-2)c)^2} \end{aligned} \quad (125)$$

#### 4.4.2 Changes in the size of the complete graph for our last example

When we change the size of the complete graph we can see for example what size would be the most effective for increasing ones PageRank. In all these examples we will use  $n_L = 10, j = 6, c = 0.85$  and  $n_G$  will vary between 1 and 50. First we note that the part above the complete graph is unaffected by the change of  $n_G$ . It is obvious however that as  $n_G$  increases the normalizing constant in the normalized PageRank will likely get larger resulting in a lower PageRank as long as it is part of a small system.

For the nodes in the complete graph except for the one thats part of the line we get the result in Fig. 8.

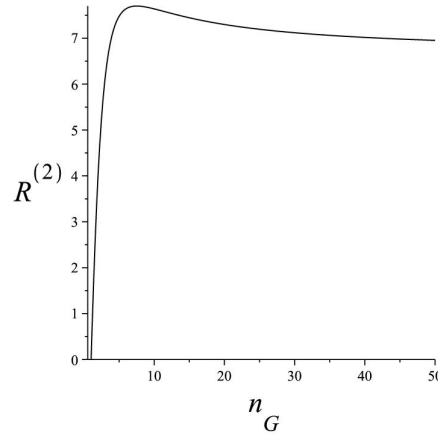


Figure 8:  $\mathbf{R}^{(2)}$  of the nodes in the complete graph not part of the line as a function of  $n_G$

Looking at the function we can see two things, first of all a larger number of nodes in the graph will increase the rank of the nodes in it. We can also see a hint that it seems to be converging towards a value as  $n_G$  gets large. Since the chance of escaping the graph decreases as  $n_G$  increases we can expect it to eventually keep nearly all of it resulting in the PageRank of all the nodes in the complete graph approaching  $1/(1 - c) \approx 6.67, c = 0.85$ .

For the node in the complete graph thats part of the line as well we get the result in Fig. 9.

Here we see something curious, the node seems to be gaining rank in the beginning while starting to fall after a while and possibly converging towards a value in the same way as the other nodes in the complete graph. The reason we get a local maximum is the fact that for a moderately large  $n_G$  we maximize the probability of  $\mathbf{R}_{L,j}^{(2)}$  getting back to itself while keeping the complete graph large enough to keep most probability for itself. Here we can see that its not always a good idea for an individual node to join a complete graph. If the node in question already have larger PageRank than the other nodes in the complete graph it actually might lose PageRank from joining it.

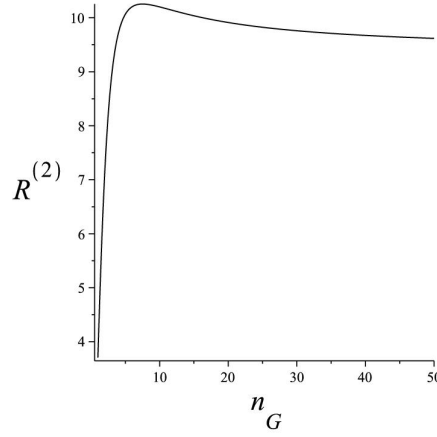


Figure 9:  $\mathbf{R}^{(2)}$  of the node in the complete graph that's part of the line as a function of  $n_G$

The result for the node below the complete graph we get the result in Fig. 10.

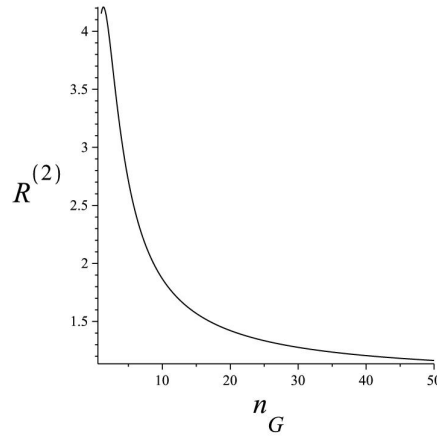


Figure 10:  $\mathbf{R}^{(2)}$  of the nodes in the line below the complete graph as a function of  $n_G$

Here we see the great loser as  $n_G$  increases. Since the chance of escaping the complete graph depends on  $\mathbf{R}_{S_G,j}^{(2)}[S_L \leftrightarrow S_G]/n_G$  as  $n_G$  increases so does this node's PageRank as well. From this we see a clear example of the effects of complete graphs on its surrounding nodes. A complete graph can be seen as a type of sink, all links to the complete graph will be used to maximum effect within the complete graph. And even worse, even if the complete graph has some nodes that point out of it their influence will be very small since the nodes in the complete graph having a large number of links the chance of escaping is low.

## 4.5 A look at the normalized PageRank for the line connected with a complete graph

Looking at the normalized PageRank in our last example with a simple line with one node being part of a complete graph we want to see how the PageRank changes as  $c$  or the relation between the size of the line or complete graph changes.

### 4.5.1 Dependence on $c$

Plotting the PageRank with  $n_G = 10, n_L = 10, j = 6$  and  $c \in [0.01, 0.99]$  we get the following results. For the node just above the complete graph  $\mathbf{R}_{S_L, i}^{(1)}[S_L \leftrightarrow S_G], i = 7$  we get the result in Fig. 11.

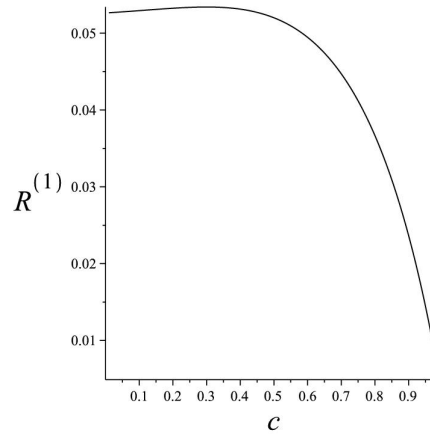


Figure 11:  $\mathbf{R}^{(1)}$  of the node above the complete graph as a function of  $c$

Here we see that the function seems to have a max at about  $c = 0.55$  after which it decreases faster the closer to  $c = 1$  it gets. We find the  $c$  which maximize the function for some other different parameters  $n_G, n_L, j, i$  in the table below. All the local max/min is calculated using the optimization tool in Maple 15.

As seen the location of the maximum seems to be moving towards the left as  $n_G$  increases and towards the right as  $n_L$  increases. In the same manner it moves towards the left as  $i$  get closer to  $n_L$ . The value of the maximum is only included out of completeness, it is natural that they decrease as either  $n_G$  or  $n_L$  increases as we in those cases get a larger number of total nodes in the system. It is interesting to note that the max seems to be going towards the right as both  $n_G, n_L$  increases as well. Looking at the node in the line being a part of the complete graph we get the result in Fig. 12.

Here we see the great "winner" as  $c$  increases. Do note the difference in the axis for the different images, since this at its lowest point is actually about the same as the highest for the node above the complete graph. The PageRank of this node is the largest when  $c$  is large, sometimes with a local maximum and sometimes not. It seems to be that as

Table 1: Maximum PageRank  $\mathbf{R}^{(1)}$  of node  $i$  "above" the complete graph depending on  $c$  for various changes in the graph where one node in a simple line is part of a complete graph.

$n_G$	$n_L$	$j$	$i$	$c_{\max}$	max
5	10	6	7	0.349	0.073
10	10	6	7	0.300	0.053
20	10	6	7	0.248	0.035
10	20	6	7	0.751	0.370
20	20	6	7	0.721	0.027
50	50	6	7	0.874	0.010
10	10	9	10	0.000	0.053
10	10	3	4	0.515	0.054
10	10	6	9	0.300	0.053

Table 2: Maximum PageRank  $\mathbf{R}^{(1)}$  of node  $j$  depending on  $c$  for various changes in the graph where node  $j$  in a simple line is part of a complete graph.

$n_G$	$n_L$	$j$	$c_{\max}$	max
5	10	6	1	0.164
10	10	6	0.894	0.099
20	10	6	0.776	0.059
10	20	6	1	0.096
20	20	6	0.929	0.056
50	50	6	0.965	0.023
10	10	9	1	0.091
10	10	3	0.893	0.107



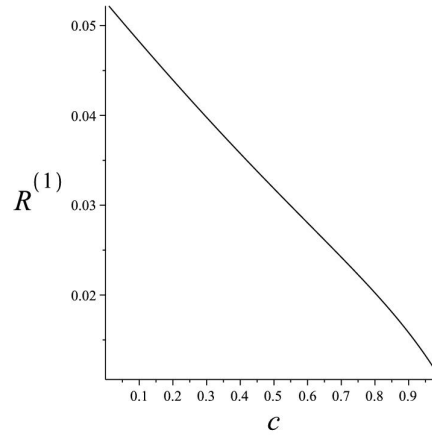


Figure 12:  $\mathbf{R}^{(1)}$  of the node in the line being a part of the complete graph as a function of  $c$

the number of nodes in the complete graph increases we are more likely to find a local maximum than not. For the node just below the complete graph we get the result in Fig. 13.

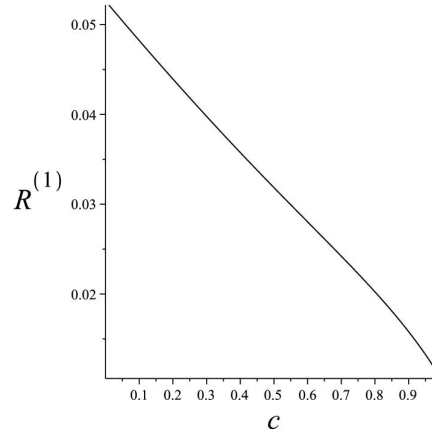


Figure 13:  $\mathbf{R}^{(1)}$  of the node below the complete graph as a function of  $c$

PageRank decreases as  $c$  increases, but compared to the nodes above the complete graph not as fast for large  $c$ . This since the PageRank of the nodes in the complete graph increase so fast for large  $c$  that even the comparatively small influence it have on the nodes out of it is enough to at least stop the extremely rapid loss of rank as for the nodes above the complete graph.

Last we got the PageRank of the other nodes in the complete graph in Fig. 14.

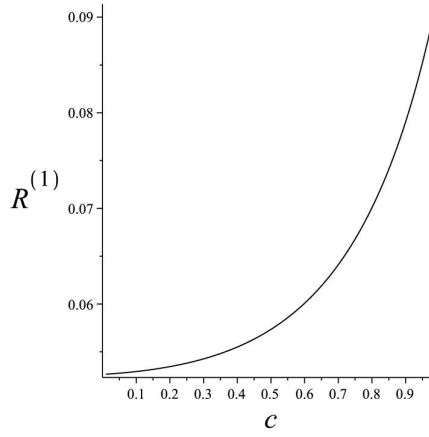


Figure 14:  $\mathbf{R}^{(1)}$  of a node in the complete graph as a function of  $c$

As with the node in both the line and complete graph, PageRank increases very fast for large  $c$ . We once again see a hint to why a too large  $c$  could be problematic, it is for large  $c$  we get the largest relative changes in PageRank between nodes. We have no min/max here, instead PageRank increases faster and faster the larger  $c$  gets.

We note that these local maximum and minimum are not always present. In these cases we have a PageRank that decreases as  $c$  increases for the whole interval. If one exists we can expect the other to as well (since we expect the rank to decrease at the end of the interval). It is hard to say anything conclusive about the location or existence of local maximum or minimum points, but we do note that they exist. There is also a large difference in how PageRank changes for different (especially large)  $c$ , we can therefore expect  $c$  to have an effect not only in the final rank and the computational aspect, but also the final ranking order of pages.

#### 4.5.2 A look at the partial derivatives with respect to $c$

Since we have the formulas for the normalized PageRank it is also possible to find the partial derivatives. Since the partial derivatives result in very large expressions (multiple pages each) they are not included here. By setting  $n_G = n_L = 10, j = 6$  we get the result after taking the partial derivative with respect to  $c$  for  $0.05 < c < 0.95$  for the node  $e_{L,7}$  above the complete graph in Fig. 15.

We see the derivative falling faster as  $c$  increases. Here as well we see the more dramatic changes in large  $c$  above about 0.8. Apart from seeing the maximum at around  $c = 0.3$  in the original function we can also see that the derivative seems to briefly increase in the beginning, reaching a maximum at about  $c = 0.1$ . For the node part of both the line and the complete graph we get the result in Fig. 16.

We can see a high derivative all the way until we get to very large  $c$  where it finally starts going down. We can clearly see the maximum at about  $c \approx 0.9$  in the original

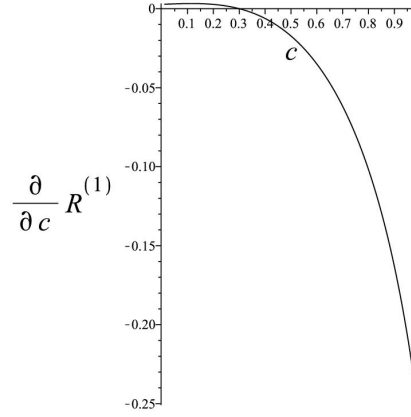


Figure 15: Partial derivative with respect to  $c$  of normalized PageRank of the node in the line above the complete graph

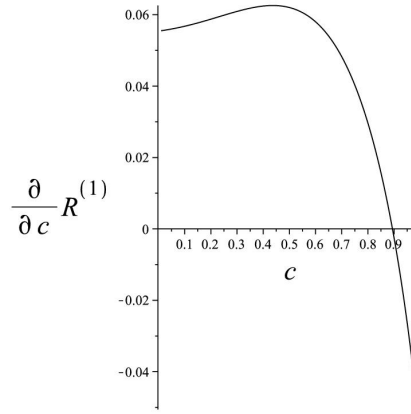


Figure 16: Partial derivative with respect to  $c$  of normalized PageRank of the node in the line part of the complete graph

function. For the node on the line below the complete graph we get the result in Fig. 17.

Although the derivative is decreasing for all  $c$ , the derivative have a local maximum at about  $c \approx 0.6$ .Worth to note is that the axis can be a little misleading, the partial derivative is in fact not that close to 0 at the local maximum. As before the largest changes are at high  $c$ . Worth to note that the derivative is decreasing for all  $c$ . For the nodes in the complete graph not part of the line we get the result found in Fig. 18

As before the largest changes are found at large  $c$ . Compared to the node part of both the complete graph and the line the derivative for the ones only in the complete graph continue to increase as  $c$  increases, however the PageRank itself is not actually

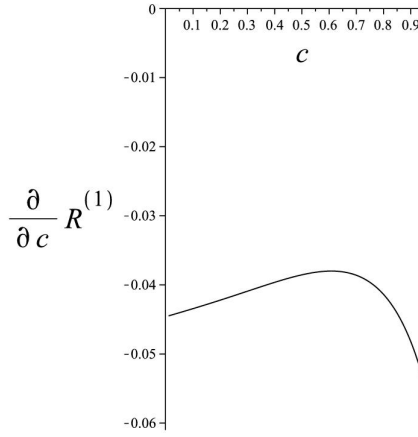


Figure 17: Partial derivative with respect to  $c$  of normalized PageRank of the node in the line below the complete graph

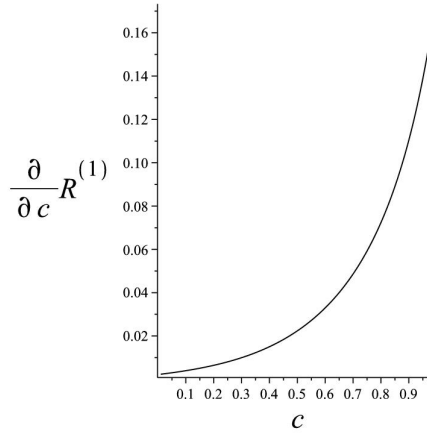


Figure 18: Partial derivative with respect to  $c$  of normalized PageRank of a node in the complete graph not part of the line

ever higher for the ones not part of the line. We have seen that although it is possible to find symbolic expressions for the PageRank and derivative for some simple graphs, as the complexity of the graph increases it becomes very hard to do. Already for these simple examples the partial derivatives a rather large and complicated expressions. Finding more general symbolic expressions for when the derivative is *zero* should be possible although problematic given the constraints and size of the problem.

## 4.6 The effect of changing the weight vector $\mathbf{V}$

From the equation system we see that the inverse matrix in the solution  $\mathbf{R}^{(2)}$  does not depend on  $\mathbf{V}$ . While there is usually the system matrix  $\mathbf{A}$  that rapidly changes making calculating PageRank in this way unpractical since we need to calculate the inverse of a huge matrix when doing changes (in the case of Internet pages). If we instead would have a mostly static system but with varying weight vector  $\mathbf{V}$  it might be useful to use this representation instead since calculating the new PageRank would then be a simple matrix-vector multiplication. We can also see that changing an element in  $\mathbf{V}$  to zero from the uniform case  $1/n_S$  has two effects. First of all we know that it would change the nodes PageRank by at least a constant amount (1 from uniform  $\mathbf{V}$ ) the rest of the probability  $c/1 - c \approx 5.67$ ,  $c = 0.85$  might be lost from other nodes in the vicinity. If the node has no outgoing links the PageRank of all other nodes will be unaffected. In the other case where the node in question is not a dangling node nor is it possible to reach any dangling nodes from it, all of it will be removed somewhere. We find the maximum that can be lost by setting a nodes weight to zero assuming a previous weight of 1 as 1 plus what we get if all nodes it link to link directly back to it and do not link to anything else as:

$$\sum_{k=0}^{\infty} c^{2k} = \frac{1}{1 - c^2} \approx 3.6, \quad c = 0.85 \quad (126)$$

In the same way doubling  $V_i$  for one node increases the PageRank of those same nodes by the same amount as they would otherwise lose had we instead set it to zero.

Especially effective it seems to simply change  $\mathbf{V}$  for nodes in a complete graph if they are believed to be cheating, since the complete graph is so effective in keeping its probability to itself changing  $\mathbf{V}$  to zero for those nodes should have a very little effect in surrounding nodes apart from possibly scaling the PageRank for all the nodes in the system with a different constant in the case of the normalized PageRank  $\mathbf{R}^{(1)}$ .

## 4.7 A comparison of normalized and non-normalized PageRank

Here we will take a short look at the difference between normalized ( $\mathbf{R}^{(1)}$ ) and non normalized ( $\mathbf{R}^{(2)}$ ) PageRank in order to get a bigger understanding of the differences between them. We already know that  $\mathbf{R}^{(2)} \propto \mathbf{R}^{(1)}$  so there will always be the same relation between the PageRank of two nodes. Here we will take a look at how the absolute difference between nodes and the two types of PageRank differ instead.

Since the PageRank is normalized to one in  $\mathbf{R}^{(1)}$  we obviously get that the PageRank will decrease as the number of nodes increases, potentially making for problems with number-representation for extremely large graphs unless it is taken into account when making the implementation. This problem is not as large a problem for  $\mathbf{R}^{(2)}$  since most nodes will have approximately the same size regardless of the size of the graph. However the possible huge relative difference between nodes is still needed to take into consideration.. We note however that with the current way to calculate  $\mathbf{R}^{(2)}$  by solving the equation system such large systems that could potentially be a problem in  $\mathbf{R}^{(1)}$  is simply too large for us to solve in a timely manner.

We also have one other main difference between the normalized and non-normalized PageRank and that is with dangling nodes and how they effect the global PageRank. In  $\mathbf{R}^{(2)}$  a dangling nodes means some of the "probability" escape the graph resulting in a lower total PageRank (but still proportional to  $\mathbf{R}^{(1)}$ ). In  $\mathbf{R}^{(1)}$  however dangling nodes can be seen as linking to all nodes and in fact behaves exactly as if they did. We illustrate the difference in a rather extreme example with a graph composed of only four dangling nodes as well as a complete graph composed of four nodes.

An image of the systems can be seen in Fig. 19 below. When computing  $\mathbf{R}^{(1)}$  of both



Figure 19: A complete graph (left) and a system made of four dangling nodes (right)

systems assuming uniform weightvector  $\mathbf{u}$  they are both obviously equal with PageRank  $\mathbf{R}^{(1)} = [1/4, 1/4, 1/4, 1/4]$ , it does not even matter what  $c$  we chose as long as it is between zero and one for convergence. However for the non normalized PageRank we get a large difference between the PageRank of the two systems where we for the complete graph get the PageRank  $\mathbf{R}_a^{(2)} = [1/1 - c, 1/1 - c, 1/1 - c, 1/1 - c]$  as seen in Sect. 4.2. However for the graph made up of only dangling nodes we get the PageRank  $\mathbf{R}_b^{(2)} = [1, 1, 1, 1]$  regardless of  $c$ . We see that while they might be proportional to each other, the non normalized version behaves differently for dangling nodes making a distinction between dangling nodes and nodes that link to all nodes (including itself which we normally do not allow). While this distinction might seem unnecessary since nodes that link to all nodes do not normally exist or similar nodes such as a node that links to all or most other nodes should either be extremely uncommon or plain do not exist as well, this might not be the case if working with smaller link structures where such a distinction might be useful. It is also this distinction that makes it possible to make comparisons of PageRank between different systems in  $\mathbf{R}^{(2)}$  while not generally possible in  $\mathbf{R}^{(1)}$ .

## 5 Conclusions

We have seen that we can solve the resulting equation system instead of using the definition directly or using the Power method. While this method is significantly slower it has made it possible to get a bigger understanding of the different roles of the link matrix  $\mathbf{A}$  and the weight vector  $\mathbf{u}$ . We have seen how PageRank changes when doing some small changes in a couple of simple systems and when connecting said systems.

For these systems we also found explicit expressions for the PageRank and in particular two ways to find these. Either by solving the equation system itself or by calculating:

$$\left( \sum_{e_i \in S, e_i \neq e_g} P(e_i \rightarrow e_g) + 1 \right) \left( \sum_{k=0}^{\infty} (P(e_g \rightarrow e_g))^k \right)$$

where  $P(e_i \rightarrow e_g)$  is the sum of probability of all paths from node  $e_i$  to node  $e_g$  and the weight vector  $\mathbf{u}$  is uniform.

Given the expressions for PageRank we looked at the results when changing some parameters. While it is hard to say anything specific, two things seem to be true overall: The most dramatic changes happens as  $c$  get large, usually somewhere where  $c > 0.8$  some nodes get dramatically larger PageRank compared to the other. We also see that complete graphs, while not gaining a larger rank if the graph is larger, it becomes a lot more reliable (as in not as effected in changes of individual nodes) in keeping its large PageRank as the structure get larger.

We saw that if using uniform  $\mathbf{V}$  it is possible to split a large system  $S$  into multiple disjoint systems  $S_1, S_2, \dots S_N$  it is possible to calculate  $\mathbf{R}^{(2)}$  for every subsystem itself and they will not differ from  $\mathbf{R}^{(1)}$  apart from a normalizing constant that is the same across all subsystems. This is a property we would like to if possible have when using the power method as well. This since it could potentially greatly reduce the work needed primary when doing updates in the system.

For the last part we looked at what happens in  $\mathbf{R}^{(2)}$  when changing the weight vector  $\mathbf{V}$ . Especially we could see some guaranteed change in the constant change and we could find an upper bound in how much total difference the change can have overall. Especially effective it seems to be in lowering the PageRank of nodes in complete graphs since they keep most of their probability to themselves.

## 6 Acknowledgments

This research was supported in part by the Swedish Research Council (621- 2007-6338), Swedish Foundation for International Cooperation in Research and Higher Education (STINT), Royal Swedish Academy of Sciences, Royal Physiographic Society in Lund and Crafoord Foundation.

## References

- [1] Adaptive methods for the computation of pagerank. *Linear Algebra and its Applications*, 386(0):51 – 65, 2004. Special Issue on the Conference on the Numerical Solution of Markov Chains 2003.
- [2] F. Andersson. Estimation of the quality of hyperlinked documents using a series formulation of pagerank. Master’s thesis, Mathematics, Centre for Mathematical sciences, Lund Institute of Technology, Lund University, May 2006:E22. LUTFMA-3132-2006.

- [3] F. Andersson and S. Silvestrov. The mathematics of internet search engines. *Acta Appl. Math.*, 104, 2008.
- [4] A. Berman and R. Plemmons. *Nonnegative Matrices in the Mathematical Sciences*. Number del 11 in Classics in Applied Mathematics.
- [5] D. Bernstein. *Matrix Mathematics*. Princeton University Press, 2005.
- [6] M. Bianchini, M. Gori, and F. Scarselli. Inside pagerank. *ACM Trans. Internet Technol.*, 5(1):92–128, Feb. 2005.
- [7] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107 – 117, 1998. Proceedings of the Seventh International World Wide Web Conference.
- [8] K. Bryan and T. Leise. The \$25, 000, 000, 000 eigenvector: The linear algebra behind google. *SIAM Review*, 48(3):569–581, 2006.
- [9] D. Dhyani, S. S. Bhowmick, and W.-K. Ng. Deriving and verifying statistical distribution of a hyperlink-based web page quality metric. *Data Knowl. Eng.*, 46(3):291–315, Sept. 2003.
- [10] C. Engström. Pagerank as a solution to a linear system, pagerank in changing systems and non-normalized versions of pagerank. Master’s thesis, Mathematics, Centre for Mathematical sciences, Lund Institute of Technology, Lund University, May 2011:E31. LUTFMA-3220-2011.
- [11] F. Gantmacher. *The Theory of Matrices*. Gantmacher.
- [12] T. Haveliwala and S. Kamvar. The second eigenvalue of the google matrix. Technical Report 2003-20, Stanford InfoLab, 2003.
- [13] H. Ishii, R. Tempo, E.-W. Bai, and F. Dabbene. Distributed randomized pagerank computation based on web aggregation. In *Decision and Control, 2009 held jointly with the 2009 28th Chinese Control Conference. CDC/CCC 2009. Proceedings of the 48th IEEE Conference on*, pages 3026–3031, 2009.
- [14] S. Kamvar and T. Haveliwala. The condition number of the pagerank problem. Technical Report 2003-36, Stanford InfoLab, June 2003.
- [15] S. D. Kamvar, M. T. Schlosser, and H. Garcia-Molina. The eigentrust algorithm for reputation management in p2p networks. In *Proceedings of the 12th international conference on World Wide Web, WWW ’03*, pages 640–651, 2003.
- [16] P. Lancaster. *Theory of Matrices*.
- [17] J. R. Norris. *Markov chains*. Cambridge University Press, New York, 2009.
- [18] R. Tobias and L. Georg. *Markovprocesser*. Univ., Lund, 2000.